Technische Universität München

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

A Bayesian analysis of rare B decays with advanced Monte Carlo methods

Frederik Beaujean

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Martin Beneke
Prüfer der Dissertation:	1. Hon.-Prof. Allen C. Caldwell, Ph.D.
	2. Univ.-Prof. Dr. Stephan Paul

Die Dissertation wurde am 08.10.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 12.11.2012 angenommen.

Zusammenfassung

Auf der Suche nach neuer Physik werden seltene Zerfälle von B-Mesonen betrachtet, die durch $b \to s$ Übergänge charakterisiert sind. Hierzu wird ein modellunabhängiger globaler Fit durchgeführt, in dem die Kopplungen bzw. Wilsonkoeffizienten $\mathcal{C}_7$, $\mathcal{C}_9$ und $\mathcal{C}_{10}$ der $\Delta B = 1$ effektiven Feldtheorie bestimmt werden. Unter der Annahme reellwertiger $\mathcal{C}_i$ werden dabei alle Operatoren, die im Standardmodell $b \to s\gamma$ und $b \to s\ell^+\ell^-$ Übergänge beschreiben, betrachtet. Von den Experimenten BaBar, Belle, CDF, CLEO und LHCb gehen insgesamt 59 Messungen von Observablen aus den Zerfällen $B \to K^*\gamma$, $B \to K^{(*)}\ell^+\ell^-$ und $B_s \to \mu^+\mu^-$ ein.

Die vorgestellte Analyse ist die erste ihrer Art, die den Bayesschen Zugang zur Wahrscheinlichkeitstheorie vollständig ausnutzt. Alle wichtigen Beiträge zur Theorieunsicherheit werden explizit mit Hilfe von Nuisanceparametern abgebildet. Auf diese Weise wird die Information aus den Messungen optimal genutzt, um gleichzeitig die Wilsonkoeffizienten und die Nuisanceparameter, insbesondere die Formfaktoren, einzuschränken. Letztere stellen die größte Quelle von Theorieunsicherheit dar.

Aus numerischer Sicht besteht die Aufgabe darin, Zufallszahlen nach der a-posteriori Wahrscheinlichkeitsverteilung P zu ziehen, um damit die marginalisierten Verteilungen der Fitparameter zu bestimmen und die Vorhersage bisher nicht gemessener Observablen per Fehlerfortpflanzung zu ermöglichen. Dies wird durch zwei Punkte erschwert. Zum einen ist der Parameterraum hochdimensional, und P hat mehrere weit entfernte Maxima sowie Entartungen. Zum anderen ist die Berechnung der Theorievorhersagen, die im Fit mit den Messdaten verglichen werden, sehr rechenaufwändig. Eine einzelne Auswertung von P benötigt ca. 1 s, insgesamt sind einige Millionen Auswertungen nötig. Population Monte Carlo (PMC) löst beide Probleme auf einmal. Hierzu wird eine Mischverteilung schrittweise an P angepasst, sodass per Importance Sampling die Zufallszahlen auf massiv-parallele Art gezogen werden können. Als hinderlich erweist sich die empfindliche Abhängigkeit von PMC auf die Initialisierung, für die P bereits relativ gut bekannt sein muss. Auf dem Weg zu einem allgemeinen, problemunabhängig funktionierenden Monte Carlo Algorithmus wird ein neue Methode entwickelt, welche die nötige Information über P automatisch und zuverlässig aus Markov-Ketten mittels hierarchischem Clustering gewinnt.

Unter Mitnahme der neuesten experimentellen Ergebnisse aus dem Jahr 2012 zeigt der Fit zwei getrennte Bereiche hoher Wahrscheinlichkeit. Neben einer dem Standardmodell ähnlichen Lösung verbleibt auch eine Lösung mit umgekehrten Vorzeichen, die wegen der näherungsweisen Invarianz aller Observablen unter $\mathcal{C}_i \to -\mathcal{C}_i$ ähnlich wahrscheinlich ist. An beiden lokalen Maxima von P werden die Messdaten gut beschrieben. Der Standardmodell-Punkt ist nahe am globalen Maximum von P. Obwohl der Fit noch große Abweichungen vom Standardmodell in $\mathcal{C}_i$ zulässt, zeigen sich dennoch keine zwingenden Hinweise auf neue Physik, da das Standardmodell als einfachere Beschreibung der Daten durch den Bayes-Faktor klar bevorzugt wird.

Es werden zwei Sätze von Vorhersagen für Observablen in der Winkelverteilung von $B \to K^*(\to K\pi)\,\ell^+\ell^-$ berechnet. Darin sind insbesondere bisher nicht gemessene, optimierte Observablen enthalten, die aufgrund ihrer geringen Formfaktor-Abhängigkeit und der Empfindlichkeit auf vom Standardmodell abweichende Wechselwirkungen für zukünftige Analysen von herausragendem Interesse sind. Zum einen werden mit einer Bayesschen Methode Standardmodell-Vorhersagen abgeleitet, die in guter Übereinstimmung mit der Literatur stehen. Des weiteren wird das verbesserte Wissen über die Nuisanceparameter aus dem Fit verwendet. Hierbei ergeben sich beträchtlich genauere Vorhersagen für Observablen, die von Formfaktoren abhängen. Sollten zukünftige Messungen von diesen Vorhersagen abweichen, wären sie als deutliche Hinweise auf neue Physik jenseits des hier betrachteten Szenarios mit Standardmodell-Operatoren zu werten.

Abstract

Searching for new physics in rare B meson decays governed by $b \to s$ transitions, we perform a model-independent global fit of the short-distance couplings $\mathcal{C}_7$, $\mathcal{C}_9$, and $\mathcal{C}_{10}$ of the ΔB=1 effective field theory. We assume the standard-model set of $b \to s\gamma$ and $b \to s\ell^+\ell^-$ operators with real-valued $\mathcal{C}_i$. A total of 59 measurements by the experiments BaBar, Belle, CDF, CLEO, and LHCb of observables in $B \to K^*\gamma$, $B \to K^{(*)}\ell^+\ell^-$, and $B_s \to \mu^+\mu^-$ decays are used in the fit. Our analysis is the first of its kind to harness the full power of the Bayesian approach to probability theory. All main sources of theory uncertainty explicitly enter the fit in the form of nuisance parameters. We make optimal use of the experimental information to simultaneously constrain the Wilson coefficients as well as hadronic form factors — the dominant theory uncertainty.

Generating samples from the posterior probability distribution to compute marginal distributions and predict observables by uncertainty propagation is a formidable numerical challenge for two reasons. First, the posterior has multiple well separated maxima and degeneracies. Second, the computation of the theory predictions is very time consuming. A single posterior evaluation requires $\mathcal{O}(1\,\mathrm{s})$, and a few million evaluations are needed. Population Monte Carlo (PMC) provides a solution to both issues; a mixture density is iteratively adapted to the posterior, and samples are drawn in a massively parallel way using importance sampling. The major shortcoming of PMC is the need for cogent knowledge of the posterior at the initial stage. In an effort towards a general black-box Monte Carlo sampling algorithm, we present a new method to extract the necessary information in a reliable and automatic manner from Markov chains with the help of hierarchical clustering.

Exploiting the latest 2012 measurements, the fit reveals a flipped-sign solution in addition to a standard-model-like solution for the couplings $\mathcal{C}_i$. The two solutions are related by approximate invariance of the observables under $\mathcal{C}_i \to -\mathcal{C}_i$. Both solutions contain about half of the posterior probability and provide a good fit to the data. The standard-model prediction is close to the global best-fit point. The Bayes factor strongly favors the simplicity of the standard model over the more complex new-physics scenario.

For future searches, we compute two sets of predictions of observables in the angular distributions of $B \to K^*(\to K\pi)\,\ell^+\ell^-$. This includes currently unmeasured optimized observables with reduced form-factor dependence and sensitivity to nonstandard interactions. In the first set, predictions within the standard model are calculated with a Bayesian approach and found to agree well with existing results. In the second set, we make use of the improved posterior knowledge of nuisance parameters to compute predictions based on the new-physics fit output. In the latter case, we observe significantly reduced theory uncertainty for all observables with form-factor dependence. Deviations from the predictions in future measurements of the predicted observables would clearly indicate new physics beyond the considered scenario in the standard-model operator basis.

Contents

1 Introduction

High energy physics has entered an exciting phase. Just a little over two years after the start of the physics program at the *large hadron collider* (LHC) in the spring of 2010, the two collaborations ATLAS [Aad+12a] and CMS [Cha+12a] presented strong evidence of the existence of a new boson. So far, little is known about this boson beyond its mass near 125 GeV and its spin (0 or 2), but these facts suggest that it is the long-sought Higgs boson, predicted nearly 50 years ago [EB64; Hig64; GHK64]. If that were true, it would be the last piece in a sequence of elementary particles that had been predicted by the *standard model* (SM) of particle physics and were subsequently observed. Examples include the b quark [Her+77], the t quark [Aba+95; Abe+95] and the τ neutrino [Kod+01].

But with the standard model apparently complete, we have not reached the end of particle physics. Over the past years, a number of experimental facts have accumulated that cannot be explained within the standard model. First, the confirmation of neutrino oscillation [Cle+98] has clearly demonstrated that neutrinos do have a small mass, yet they are treated as massless particles in the standard model. A new state, a heavy sterile right-handed neutrino could explain the small masses of the left-handed SM neutrinos via the "see-saw" mechanism. Second, galaxy rotation curves, galaxy (cluster) formation, and the mismatch between visible matter and the total matter density inferred from the cosmic microwave background hint at a new form of matter that interacts only weakly and gravitationally, the cold *dark matter*. No particle with the right characteristics exists within the SM. Third, the observation of the accelerated expansion of the universe provides evidence of a substance with an exotic equation of state that is held responsible for about 75 % of the energy density in the universe — *dark energy*. Fourth, considering charge (C) and parity (P) transformations, the amount of CP violation in the SM is too small to quantitatively account for the observed discrepancy between matter and antimatter in nature. It appears that some form of new physics (NP) interaction is needed at very high energies, that is in the early universe, to explain baryogenesis.

In addition, there are a number of open questions in the standard model that call for a profound explanation by a more fundamental theory of particle physics. For example, why are there exactly three generations of quarks and leptons? Or why is the θ term that breaks CP symmetry in *quantum chromodynamics* (QCD), the gauge theory of strong interactions, so small? A dynamical solution for this *strong* CP problem is provided by a hypothetical new particle, the *axion*. And why is the Higgs mass so much smaller than the Planck mass of $\mathcal{O}\left(10^{18}\,\mathrm{GeV}\right)$? One popular solution to this hierarchy problem is to assume *supersymmetry*, thereby predicting *superpartners* for the SM particles. An open area of research is to unite the SM description of the strong, weak, and electromagnetic force with the force of gravity to obtain a theory of quantum gravity. At present, the most promising candidate of such a theory of everything is *super string theory* that predicts a plethora of new states.

Most probably there is not one simple theory that solves all of the puzzles introduced above, but there certainly is a recurring theme — to require *new* particles. Therefore, the overarching goal of this work is to search for the presence of new states in order to make progress in answering the fundamental questions. There are two complementary approaches at current collider experiments. On the one hand, higher beam energies allow the *direct* production of heavy new particles, visible as resonances in invariant mass spectra. On the other hand, those particles also leave measurable traces if they are off shell; i.e. when they *indirectly* modify reactions allowed in the SM via quantum corrections. In either case, the goal is to observe a discrepancy between SM predictions and measurements.

In this work, we concentrate on the indirect searches, and we study reactions involving the transition of a b quark into an s quark, where large data sets are available. In the standard model, this transition is mediated by the weak force via a *flavor changing neutral current* (FCNC), hence there is no SM contribution at tree level that could mask NP effects. The branching ratios involving $b \to s$ are small due to loop as well as CKM suppression, yet large enough to be observed, as opposed to the similar $t \to c$ transition that is further suppressed by the GIM mechanism (see Section 5.1.2). *Rare decays* of B mesons involving $b \to s$ transitions occur with SM branching ratios ranging from 10^{-4} (radiative $B \to K^*\gamma$) over 10^{-7} (semileptonic $B \to K^{(*)}\ell^+\ell^-, \ell = e, \mu$) to 10^{-9} (leptonic $B_s \to \mu^+\mu^-$); see Section 6.3 for more details.

These decay modes are of special interest for multiple reasons. First, SM extensions like supersymmetry predict significantly enhanced branching fractions for certain parameter values. Second, experiments were only recently able to observe these rare decays, and the precision will improve dramatically within this decade. Third, the theoretical description through an *effective field theory* (EFT) provides the ideal tool to perform a model-independent separation of the short-distance, high-energy scales where we hope to see signs of NP and the long-distance, low-energy scales where we have to deal with nonperturbative QCD effects, the major source of theory uncertainty.

Chances are that signs of new physics in B decays are rather tenuous, and there may be no single observation showing a significant deviation from SM expectations. Therefore, our goal in this work is to construct a *global fit* of the effective theory of $b \to s$ transitions to a large number of experimental observations of rare B decays and to search for a mismatch. In particular, we want to fit the effective couplings, or *Wilson coefficients*, that contain the short-distance effects.

The earliest measurement we include is the determination of the $B \to K^*\gamma$ branching ratio by the CLEO collaboration at the Cornell electron storage ring [Coa+00] from 1999. Around that time, the two first-generation B factories with e^+e^- colliders, PEP-II with the BaBar detector in the USA [Aub+02], and KEKB with the Belle detector [Aba+02] in Japan, started operating. Both BaBar and Belle discovered CP violation within the $B - \bar{B}$ system [Abe+01; Aub+01], and measured the basic observables like branching ratios, forward-backward asymmetries, and longitudinal polarizations in $B \to K^*\gamma$ and $B \to K^{(*)}\ell^+\ell^-$.

B factories provide a clean environment and a well defined initial energy, so certain reactions like $B \to K^*\gamma$ are best observed there, but at the expense of a small $B\bar{B}$ production cross section of about 0.001 μb. On the contrary, reactions like $B \to K^*\ell^+\ell^-$ with muons in the final state can be well observed at hadron colliders where the production cross section is a factor of 100 000 larger. Therefore, we also use data from the general-purpose CDF experiment at the Tevatron and from LHCb, the LHC experiment with a

B-physics focus. LHCb is the only experiment that currently collects data.

The first LHCb results with 900 $B \to K^* \ell^+ \ell^-$ events [Par12], based on the 2011 data at a center-of-mass energy of 7 TeV and integrated luminosity of 1 fb^{-1}, have the smallest statistical uncertainty for this process of all experiments. Due to the increase of energy and luminosity in the 2012 run, LHCb expects a total of $\mathcal{O}(3000)$ events before LHC is shut down for maintenance and upgrades in early 2013. With this number, LHCb will have the highest accuracy in *exclusive* decays with a meson (K or K^*) and two charged leptons in the final state until the advent of the super flavor factory Belle II after about 2015. In conclusion, the near future promises significantly more accurate results on the experimental side.

With more precise input from the experiments, it is of utmost importance to improve the theory uncertainty accordingly in order to discover signs of new physics. Low-energy hadronic physics — form factors in particular — are the major source of theory uncertainty for exclusive decays; our strategy is to simultaneously fit the form factors *and* the Wilson coefficients, as the data constrain both. To this end, we model all uncertainties — CKM parameters, quark masses, form factors, missing subleading corrections — explicitly with the help of 28 nuisance parameters. Cogent prior knowledge of these parameters is available; the natural language to coherently include it in the fit is Bayesian probability theory. Another advantage of the Bayesian approach is the ability to phrase, and quantitatively answer, the following question: given the data, which model is more probable? Is it the SM, or an extension involving NP? This is the *central* question we seek to answer in this thesis.

The treatment of theory uncertainty through nuisance parameters sets us apart from previous analyses [BHD10; BHD11b; Des+11; APS12; Bob+12; AS12], and allows us to make complete and consistent use of the information available. By choosing the Bayesian framework, we can directly use the posterior knowledge of the nuisance parameters to make improved predictions for observables that, on the one hand, are sensitive to the Wilson coefficients, but, on the other hand, were not measured yet. However, the more detailed modeling comes at a price; obtaining the marginal distributions of a complicated 30 dimensional posterior density with multiple maxima and degeneracies poses a tough numerical problem; standard *Markov chain Monte Carlo* (MCMC) techniques fail to give proper results, because individual chains are trapped in local modes. To make matters worse, a single evaluation of the posterior requires approximately 0.3 s on a state-of-the-art 3.4 GHz CPU, and we need a few million evaluations in the course of the global fit. The most naïve — and most inefficient — implementation would then require more than 10 days to complete.

Our choice is to merge MCMC with adaptive importance sampling, or *population Monte Carlo* (PMC), a new technique [Cap+08] that promises to cure the major shortcomings of MCMC. In addition, importance sampling provides an estimate of the posterior normalization needed for model comparison along with the marginal distributions at no extra cost, and it is well suited to run on a massively parallel computing architecture. However, PMC comes with its own set of problems; most prominently, it is highly dependent on the initialization — we need to know the location and shape of the support of the posterior before PMC can exert its full power.

The major unpublished contribution of this work is the description of a new robust algorithm to perform that initialization. No knowledge beyond the prior information is required from the user, and all relevant features of the posterior are explored by first running multiple Markov chains. The information from the ensemble of chains is then brought into a form suitable for PMC with the help of hierarchical clustering. The good performance in the global fit demonstrates the power of this combination of MCMC and PMC, which is of general use and not tied to any specifics of B physics.

The contents of this thesis are organized as follows. The basics of Bayesian probability theory are established in Chapter 2. In Chapter 3, we review two fundamental Monte Carlo techniques, MCMC and (adaptive) importance sampling. Next, we present our new algorithm that combines the two techniques with numerous examples to highlight its strengths and potential pitfalls in Chapter 4. A concise summary of the theoretical background of rare B decays is given in Chapter 5, followed by a more detailed discussion of the observables used in the fit in Chapter 6. In Chapter 7 we show the main results of the global fit. The Appendices A – C contain details on the input numbers, priors, tables with predictions of new observables, and goodness of fit. Finally, we present background material on Monte Carlo and other numerical methods in Appendices D – F.

2 Bayesian probability theory

The main objective of this short chapter is to introduce the foundations of probability theory and to establish the notation needed to construct the global fit in Chapter 7. We follow the reasoning of the excellent book by Jaynes [JB03]. Other useful text books on the Bayesian approach include [DAg03; Ken+04; SS06].

2.1 Axioms and basic definitions

The scope of Bayesian probability theory is extremely wide: whenever one seeks to perform logical reasoning involving uncertainty, Bayesian probability applies. The rules of reasoning are based on the following three axioms:

Axiom I *The* plausibility, *or* degree of belief, *associated with a logical proposition A, is described by a real value, $P(A)$, the* probability *of A.*

If A represents a discrete set of mutually exclusive propositions, $P(A)$ is a real-valued function taking values in $[0, 1]$.

Axiom II *Qualitative correspondence with common sense.*

A small change of information should result in a small change of the degree of belief in a definite direction.

Axiom III *Consistency:*

a) For fixed information, different ways of reasoning must produce the identical result.

b) All available, relevant information is used in the reasoning, and no piece of information is arbitrarily ignored.

c) Equivalent states of knowledge yield equivalent degrees of belief.

For practical calculations, it is more convenient to use the axioms proposed by Kolmogorov [Kol33], formulated in the language of abstract set theory. Let S be a set, called the *sample space*, and let A, B denote subsets of S.

Axiom 1 *The probability is described by a real number:*

$$P(A) \geq 0 \, .$$

Axiom 2 *Sum rule:*

$$A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B) \, .$$

Axiom 3 *Normalization:*

$$P(S) = 1 \, .$$

Axiom 4 *Continuity at zero:*

$$A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots \to \varnothing \Rightarrow \lim_{n\to\infty} P(A_n) = 0 \,.$$

The above axioms can be derived from the basic reasoning Axioms I – III; cf. [JB03, App. A.1]. The *conditional* probability is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \,. \tag{2.1}$$

Given a partition of S into mutually disjoint subsets, $\bigcup_i A_i = S$, Kolmogorov's axioms and (2.1) immediately yield the *law of total probability*

$$P(B) = \sum_i P(B|A_i)P(A_i) \,. \tag{2.2}$$

Plugging (2.2) into (2.1) and using the commutativity $A \cap B = B \cap A$, we obtain the discrete formulation of *Bayes' theorem* [BB58]

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \,. \tag{2.3}$$

For propositions indexed by a real rather than an integer number, we use the same symbol $P(\cdot)$ to denote the probability density function (PDF). The equations (2.1) – (2.3) are then modified with summation replaced by integration. For example, consider a *statistical model* M — a set of assumptions used to make predictions for an experiment — with one real parameter θ, and observations denoted by D. The continuous version of Bayes' theorem, derived in [Har83], is

$$\boxed{P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta,M)}{Z} \,.} \tag{2.4}$$

Each term in Bayes' theorem, the central equation of probability theory, is referred to by a specific identifier. $P(\theta,M)$ is the *prior density*, $P(D|\theta,M)$ is called the *probability of the data* when treated as a function of D, and known as the *likelihood* when considering the dependence on θ, for fixed D. The model-dependent normalization constant Z is known as the *evidence* or *marginal likelihood*:

$$Z = \int \mathrm{d}\theta \, P(D|\theta,M)P(\theta,M) \,. \tag{2.5}$$

Finally, the left-hand side of (2.4), $P(\theta|D,M)$, is the *posterior density*. Prior and posterior ("density" is usually omitted) represent the state of knowledge of the parameter θ before and after seeing the data. Note that θ appears on opposite sides of " | " in $P(D|\theta,M)$ and $P(\theta|D,M)$. Bayes' theorem is therefore known also as the theorem of *inverse probability*.

For one dimensional problems, we define the monotonously increasing cumulative distribution function (CDF) as the integral of the corresponding PDF:

$$F(a) = \int_{-\infty}^{a} \mathrm{d}\theta \, P(\theta) \,. \tag{2.6}$$

The following quantities are often used to characterize a 1D PDF:
Expectation value or *mean*

$$E_P[\theta] \equiv \int \mathrm{d}\theta\, P(\theta)\, \theta\ , \tag{2.7}$$

mode

$$\theta^* \equiv \arg\max_\theta P(\theta)\ , \tag{2.8}$$

and *variance*

$$V_P[\theta] \equiv \int \mathrm{d}\theta\, P(\theta)(\theta - E_P[\theta])^2\ . \tag{2.9}$$

When the model contains more than one parameter, we will use the notation $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$. The *covariance* between θ_1 and θ_2 is defined as

$$\mathrm{Cov}_P[\theta_1, \theta_2] \equiv E_P[(\theta_1 - E_P[\theta_1])\,(\theta_2 - E_P[\theta_2])] = E_P[\theta_1\theta_2] - E_P[\theta_1]E_P[\theta_2]\ , \tag{2.10}$$

and the *correlation coefficient* is

$$\rho_P[\theta_1, \theta_2] \equiv \frac{\mathrm{Cov}_P[\theta_1, \theta_2]}{\sqrt{V_P[\theta_1]V_P[\theta_2]}}\ , \qquad \rho_P \in [-1, 1]\ . \tag{2.11}$$

Two parameters are *independent* if their joint distribution factorizes

$$P(\theta_1, \theta_2) = P(\theta_1)P(\theta_2) \Rightarrow \mathrm{Cov}_P[\theta_1, \theta_2] = \rho_P[\theta_1, \theta_2] = 0\ . \tag{2.12}$$

Suppose the set of parameters is partitioned into the *parameters of interest*, $\boldsymbol{\theta}$, and the *nuisance parameters*, $\boldsymbol{\nu}$. At the fundamental level of Bayes' theorem, there is no distinction between $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$. However, the goal of the analysis is to extract the posterior of $\boldsymbol{\theta}$, while $\boldsymbol{\nu}$ is only needed at an intermediate stage; for example in order to correctly model the measurement process of D. From the joint posterior $P(\boldsymbol{\theta}, \boldsymbol{\nu}|D)$, we compute the *marginalized* posterior and remove $\boldsymbol{\nu}$ by integration:

$$P(\boldsymbol{\theta}|D) = \int \mathrm{d}\boldsymbol{\nu}\, P(\boldsymbol{\theta}, \boldsymbol{\nu}|D)\ . \tag{2.13}$$

If there is only a single model under consideration, and no potential for confusion, the model label M is implied and usually omitted from the equations. But suppose that there are two competing models, M_1, M_2, with parameters $\boldsymbol{\theta}_{1,2}$, that quantitatively predict the outcome D of an experiment. The task is to find the model with the higher degree of belief. Using Bayes' theorem, the *posterior odds* of the models are easily found as

$$\frac{P(M_1|D)}{P(M_2|D)} = B_{12} \cdot \frac{P(M_1)}{P(M_2)}\ , \tag{2.14}$$

where the *Bayes factor* of M_1 versus M_2, B_{12}, is just the ratio of the evidences

$$B_{12} = \frac{P(D|M_1)}{P(D|M_2)} = \frac{Z_1}{Z_2} = \frac{\int \mathrm{d}\boldsymbol{\theta}_1\, P(D|\boldsymbol{\theta}_1, M_1)P(\boldsymbol{\theta}_1, M_1)}{\int \mathrm{d}\boldsymbol{\theta}_2\, P(D|\boldsymbol{\theta}_2, M_2)P(\boldsymbol{\theta}_2, M_2)} \tag{2.15}$$

The *prior odds* $P(M_1)/P(M_2)$ represent the relative degree of belief in the models, independent of the data. The Bayes factor quantifies the relative shift of degree of belief induced by the data. In general, $\dim \boldsymbol{\theta}_1 \neq \dim \boldsymbol{\theta}_2$, and without loss of generality let $\dim \boldsymbol{\theta}_1 < \dim \boldsymbol{\theta}_2$. The Bayes factor automatically penalizes M_2 for its larger complexity, as the prior mass is spread out over a higher-dimensional volume. However, this can be compensated if the likelihood $P(D|\boldsymbol{\theta}_2, M_2)$ is significantly higher in regions of reasonably high prior density; i.e. the Bayes factor implements Occam's razor [1]: the simplest model that describes the observations is preferred.

In the Bayesian approach, there is, however, no straightforward answer to the following question: if there is only one model at hand, how to decide if that model is sufficient to explain the data, or if the search for a better model needs to continue? The standard procedure to tackle this problem of evaluating the *goodness of fit* is explained in Appendix C.

2.2 Uncertainty propagation

Suppose the random variable A is a function of $\boldsymbol{\theta}$, $A = f(\boldsymbol{\theta})$, and we know the distribution of $\boldsymbol{\theta}$, $\boldsymbol{\theta} \sim P(\boldsymbol{\theta})$. Then what is the distribution of A, $P(A)$? Using the law of total probability (2.2), we find the fundamental equation

$$P(A) = \int \mathrm{d}\boldsymbol{\theta}\, P(A, \boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{\theta}\, P(A|\boldsymbol{\theta}) P(\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{\theta}\, \delta(A - f(\boldsymbol{\theta})) P(\boldsymbol{\theta}) \,, \tag{2.16}$$

with the Dirac δ distribution. In 1D with $f(\theta)$ invertible, we obtain the usual rule for a change of variables as a special case

$$P(A) = P(\theta(A)) \left| \frac{\mathrm{d}\theta}{\mathrm{d}A} \right| \,. \tag{2.17}$$

In most cases, (2.16) does not have an analytical solution. But if a set of identically distributed samples $\{\boldsymbol{\theta}^i\}$ from $P(\boldsymbol{\theta})$ is available, then we can approximate the integral by calculating the set of samples $\{A^i : A^i = f(\boldsymbol{\theta}^i)\}$ to obtain draws from $P(A)$. The approximation converges because (2.16) is a special case of the fundamental Monte Carlo principle (3.1). Here, we use the superscript i instead of a subscript to highlight that $\boldsymbol{\theta}^i$ is a draw from a distribution.

2.3 Priors

A crucial part in the Bayesian approach is to specify the state of knowledge before new data is taken into account; i.e., the priors need to be specified. We can distinguish the following cases.

Posterior The posterior of a previous analysis, $P_1(\boldsymbol{\theta}|D_1)$ is reused as a prior $P_1(\boldsymbol{\theta})$ for the next analysis:

$$P_2(\boldsymbol{\theta}|D_2) \propto P(D_2|\boldsymbol{\theta}) P_1(\boldsymbol{\theta}) \,. \tag{2.18}$$

[1] *Numquam ponenda est pluralitas sine necessitate* – Plurality [model complexity] must never be assumed without necessity, William of Occam.

Care needs to be taken that there is no double use of the data In other words, the data distribution factorizes, and we can simply define a product likelihood to analyze all data at once with the help of Bayes theorem:

$$P_2(\boldsymbol{\theta}|D_2) \propto P(D_2|\boldsymbol{\theta})P(D_1|\boldsymbol{\theta})P_0(\boldsymbol{\theta}) \ . \tag{2.19}$$

It does not matter whether D_1 or D_2 is analyzed first, in agreement with the consistency requirement of Axiom III. However, it still remains to define the original prior $P_0(\boldsymbol{\theta})$.

Indifference For a finite set of alternatives $\{A_i : i = 1 \dots N\}$, it is straightforward to define a prior expressing complete ignorance. If no A_i is favored a-priori, then each A_i is equally probable, and the normalized probability is

$$P(A_i) = \frac{1}{N}, \ i = 1 \dots N \ . \tag{2.20}$$

This is the famous Laplace *principle of indifference* [Lap20].

Symmetry In the case of continuous parameters, the concept of indifference is less well defined. Let us consider a measurement of a distance x, described by a *normal*, or *Gaussian*, distribution

$$P(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ . \tag{2.21}$$

$\mathcal{N}(\cdot|\mu,\sigma^2)$ has its mean and mode at μ, and its variance is σ^2, where σ is the *standard deviation*q. The standard normal distribution is defined $\mathcal{N}(\cdot|\mu = 0, \sigma^2 = 1)$. μ is the *location parameter*, as $\mathcal{N}(x|\mu,\sigma^2)$ is a function of $(x-\mu)$, and σ is the *scale parameter*

$$\mathcal{N}(x|0,\sigma^2) = \mathcal{N}\left(\frac{x}{\sigma}\,|\,0,1\right) \ . \tag{2.22}$$

Suppose we want to infer only the value of μ from the experiment, and assume σ known. Demanding that it be irrelevant where the origin of the coordinate system is, we are led to choose a *flat*, or *uniform*, prior: $P(\mu) = \text{const}$. On the other hand, if we are interested in σ only, and we require that it be unimportant whether we measure distances in meters or feet, then the prior choice is $P(\sigma) = 1/\sigma$, the classic *Jeffreys* rule [Jef39, Ch. 3]. Note that without additional information on the allowed ranges, both priors are not normalized, they are *improper* priors. In this and many other examples where improper priors are used, it can be shown that the resulting posterior is proper [Ber05], and ultimately only the posterior matters. The bottom line is this: the nature of the parameter in the model leads to a different prior choice.

We want to stress a subtle point with flat priors: upon transforming to another parameter, say from μ to $\nu = 1/\mu$, the densities transform as (2.17). Hence a flat prior in μ does *not* yield a flat prior in ν. While there often is a standard way to parametrize a statistical model, there is no principle to prohibit the use of an alternative set of parameters. The task is thus to define an algorithm that yields a definite prior starting from *complete ignorance*, while preserving coordinate invariance. There has been considerable debate over the last two centuries about whether complete ignorance is ill defined, but we take a stand and follow Jaynes [JB03, p.373]:

> "Just as zero is the natural starting point in adding a column of numbers, the natural starting point in translating a number of pieces of prior information is the state of complete ignorance."

Continuing the analogy, once numbers are put in, we cannot stay in the state of complete ignorance. And it would not be desirable to do so, because one cannot perform a statistical analysis without making assumptions; in fact, Axiom III demands that all information be used. Priors provide the means to transparently state what information is used; and Bayes' theorem is the rule to learn from data. It often happens that scientists disagree about what is known, and hence what the prior ought to be; even a single person may be undecided, and want to explore the effect of different priors as part of a sensitivity study. Thus it has been suggested to define *consensus* priors to investigate, for example, the existence of the SM Higgs boson [Cal11]. The data would be analyzed with different sets of priors, corresponding to say the optimists' and pessimists' view on the current status on the Higgs. It seems noncontroversial to state that, although the data are fixed, there is not only one way to interpret them; consensus priors are a straightforward way to acknowledge this insight, and they would be of great use to avoid misinterpretation. But a large fraction of the experimental high-energy physics community adheres to the tenet that there can be only one published answer, and thus there usually *is* only one answer given.

Reference Bernardo has formalized the notion of ignorance in his concept of the *reference prior* [Ber79; BB92; Ber05]. The main idea is to find a prior density that minimizes the expected impact on the posterior for a fixed likelihood. A reference prior is often desired when a collaboration seeks to communicate the outcome of an experiment in an "objective" fashion, and there is no consensus within the community on the prior state of knowledge. Here, objective means that, while the reference prior may not represent the state of knowledge of *any* single person in the collaboration, it is based only on the statistical model encoded in the likelihood, and the class of candidate priors. Note however that prior knowledge can be included transparently by restricting the class of candidate priors. The reference prior reduces to the Jeffreys prior in the 1D case. In addition, the reference prior enjoys the following properties:

1. The reference prior can be, at least in principle, derived for any statistical problem. The resulting reference posterior is proper for a sufficiently large data set D.
2. For any one-to-one change of variables $\boldsymbol{\nu}(\boldsymbol{\theta})$, the reference prior adjusts in such a way that the reference posterior $P(\boldsymbol{\nu}|D)$ is properly related to the posterior $P(\boldsymbol{\theta}|D)$ as

$$P(\boldsymbol{\nu}|D) = P(\boldsymbol{\theta}|D) \left| \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\boldsymbol{\nu}} \right| . \tag{2.23}$$

   Hence the choice of parametrization does not affect the posterior inference.

3. For repeated sampling under the assumed model, the posterior will concentrate on a region of parameter space that contains the true value of $\boldsymbol{\theta}$.

Maximum entropy The principle of *maximum entropy* (MAXENT) in probability theory was put forth most ardently by Jaynes in [JB03]. It is not to be confused with the entropy of thermodynamics. Though there is an intimate relation between the fields, the reason the term "entropy" is used in probability theory is due to its usage in the classic masterpiece by Shannon [Sha48]. MAXENT appears to be *the* answer to include prior information as constraints into discrete (prior) probability distributions $\{P_i : i = 1 \dots N\}$, where $N = \infty$ is included. The (Shannon) entropy is

$$H(P_1, \dots, P_N) = -\sum_{i=1}^{N} P_i \log(P_i) \ . \tag{2.24}$$

For example, by using Gibb's inequality

$$-\sum_{i=1}^{N} P_i \log(P_i) \leq -\sum_{i=1}^{N} P_i \log(q_i) \tag{2.25}$$

which holds for any two distributions P and q, it is easy to verify that the uniform distribution, $P_i = 1/N$, maximizes H if the only constraint is normalization, $\sum_{i=1}^{N} P_i = 1$. For continuous problems, MAXENT proves very useful, though it is not universally applicable. We will only consider a few very important examples of practical use in the global fit. Suppose we want to assign a probability distribution to θ, and we only know its mean value, μ, and the magnitude of the variation of θ, given by the variance σ^2. Then the MAXENT distribution is uniquely determined as

$$P(\theta|\mu, \sigma^2) = \mathcal{N}\left(\theta|\mu, \sigma^2\right) \ . \tag{2.26}$$

Similarly, the MAXENT distribution among all distributions with finite support $[a, b]$ (no constraint on mean and variance) is the uniform distribution:

$$\mathcal{U}_{[a,b]}(\theta) = \frac{1}{b-a} \ . \tag{2.27}$$

It is a fact that the posterior is prior dependent; that is the content of Bayes' theorem (2.4). Asymptotically, as more data is added, the likelihood, and correspondingly the posterior, converge to a (multivariate) normal distribution centering on the "true" parameter value under mild regularity conditions [Cra99; Wal69]. In that case, the prior is negligible. For a realistic problem, there never is an infinite amount of data, there may exist several posterior modes etc., and the convergence to the normal is typically quite slow. It is thus up to the researcher to investigate the effect of different, reasonable priors, and to judge whether that effect is deemed significant. If so, conclusions based on the posterior should be taken with a grain of salt, and more data needs to be collected. It is a feature of the Bayesian approach that it automatically reveals this need for a more substantial data basis.

In practice, it is hard, and often even impossible, to follow the above guidelines to elicit the (reference) prior, as there are many parameters, and the dependence of the likelihood on them is often hidden inside a computer program. Therefore, one often resorts to a convenient prior that is sufficiently diffuse in the region of the likelihood maximum, in the expectation that the prior tails are cut off by a sharply falling likelihood, and that the details around the posterior mode are dominated by the data. As explained in detail in Appendix A.1, we benefit greatly from the MAXENT principle and from posteriors of other analyses to assign priors to the nuisance parameters in the global fit.

3 Monte Carlo sampling

We begin with the *fundamental Monte Carlo* principle. For generality, we choose the notation x to denote a random variable in this chapter. If we wish to emphasize the multidimensional character, then $x \to \boldsymbol{x}$. In case we are interested in the parameter(s) of a model M, we identify $x = \theta$ as in Chapter 2. Suppose a probability density $P(x)$, often called the *target* density, and an arbitrary function $f(x)$ with finite expectation value under P

$$E_P[f] = \int \mathrm{d}x\, P(x) f(x) < \infty \,. \tag{3.1}$$

Then a set of draws $\{x^i : i = 1 \dots N\}$ from the density P is enough to estimate the expectation value. Specifically, the integral (3.1) can be replaced by the estimator (distinguished by the symbol $\widehat{\;}$)

$$\boxed{\widehat{E_P[f]} \approx \frac{1}{N} \sum_{i=1}^{N} f(x^i),\; x \sim P\,.} \tag{3.2}$$

As $N \to \infty$, the estimate converges almost surely at a rate $\propto 1/\sqrt{N}$ [RC04, Ch. 3.2] by the strong law of large numbers if $\int \mathrm{d}x\, P(x) f^2(x) < \infty$.

How does (3.2) relate to Bayesian inference? Upon applying Bayes' theorem to real-life problems, one quickly encounters integrals of the form (3.1) that cannot be computed analytically, hence one has to resort to numerical techniques. In low dimensions, say $d \leq 2$, quadrature and other grid-based methods are fast and accurate, but as d increases, these methods generically suffer from the *curse of dimensionality*. The number of function evaluations grows exponentially as $\mathcal{O}\left(m^d\right)$, where m is the number of grid points in one dimension. Though less accurate in few dimensions, Monte Carlo — i.e., random-number based — methods are the first choice in $d \gtrsim 3$ because the computational complexity is (at least in principle) independent of d.

Which function f is of interest to us? For example when integrating over all but the first dimension of $\boldsymbol{x}$, the marginal posterior probability (cf. (2.13)) that x_1 is in $[a,b]$ is given by

$$P(a \leq x_1 \leq b|D) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{x_1 \in [a,b]}\left(\boldsymbol{x}^i\right)\,, \tag{3.3}$$

with the *indicator function*

$$\mathbf{1}_{x_1 \in [a,b]}\left(\boldsymbol{x}\right) = \begin{cases} 1, x_1 \in [a,b] \\ 0, \text{else}\,. \end{cases} \tag{3.4}$$

Equation (3.3) follows immediately from the fundamental equation (3.2) with $f(\boldsymbol{x}) = \mathbf{1}_{x_1 \in [a,b]}(\boldsymbol{x})$. The major simplification arises as we perform the integral over $d-1$ dimensions simply by ignoring these dimensions in the indicator function. If the parameter range of x_1 is partitioned into bins, then (3.3) holds in every bin, and defines the histogram approximation to $P(x_1|D)$. In exact analogy, the 2D histogram approximation is computed from the samples. For understanding and presenting the results of Bayesian parameter inference, the set of 1D and 2D marginal distributions is the primary goal. Given samples from the full posterior, we have immediate access to *all* marginal distributions at once; i.e., there is no need for separate integration to obtain for example $P(x_1|D)$ and $P(x_2|D)$. This is a major benefit of the Monte Carlo method in conducting Bayesian inference.

In order to compute the evidence (2.5) for Bayesian model comparison, however, it is necessary to integrate over *all* d dimensions. Note that the posterior samples alone do not yield this information, as the precise value of the evidence, or normalization constant, is irrelevant for their generation. Thus, extra information is required, and an algorithm that produces Z along with the marginal distributions is preferred.

In this chapter, we will discuss the two classic algorithms to generate samples from an arbitrary distribution: MCMC [Met+53; Has70] and importance sampling [NMU51]. Remarkably, those ideas were conceived at the very beginning of the computer age around 1950, and Nicholas Metropolis is a coauthor of *both* original articles. There is a powerful third alternative — nested sampling — that has been developed in the new millennium [Ski06; FHB09], but for brevity we will not detail it here. All three basic ideas are continuously refined for higher efficiency, with numerous variants existing for specific problems; cf. [RC04] for a comprehensive overview of Monte Carlo methods until 2004.

Of the three, MCMC is the most general, as it naturally deals with discrete and continuous problems, but it cannot provide the evidence without additional effort [GM98; CJ01]. Both importance and nested sampling yield posterior draws and the evidence in one run. All three algorithms can be parallelized to a certain degree, but to the best of our knowledge, importance sampling is the only one that can be run *massively parallel*; i.e., using hundred or even thousands of computing cores.

3.1 Markov chains and the Metropolis-Hastings algorithm

The following review is based on the lectures held by Prof. Caldwell at the Technische Universität München in the years 2006 – 2012 [Cal10] and a similar lecture given by the author at the Universidad de Costa Rica [Bea09].

Informally, a Markov process [Mar06; RC04] is a random-number generating process without memory. Its output, a sequence of *states*, constitutes the Markov *chain*. The output at (discrete) time $t_0 + 1$ is defined entirely by the state of the chain at time t_0 and a (fixed) transition kernel, but unaffected by the history of the chain at $t < t_0$. One of the main features of Markov chains is that they easily generalize to many dimensions. We can view a Markov process as a procedure that converts a sequence of independent, uniformly distributed random numbers into a chain of dependent random numbers from an arbitrary distribution, see Fig. 3.1.

In this section, we will define Markov chains more rigorously, and show how a Markov chain transition kernel is constructed that yields correlated samples from a desired target density as the chain grows (infinitely) long. After a description of an adaptive version that tunes itself, we consider a realistic example of a multimodal distribution that resembles the problem of the global fit, where the standard MCMC approach is highly inefficient. In Appendix D, we present some MCMC extensions that we developed to deal with the multimodality. Our recommended solution, however, is

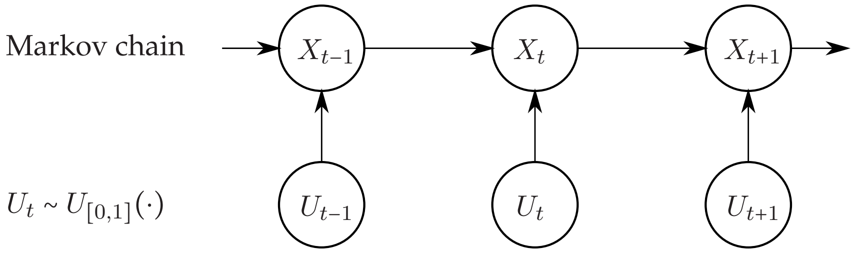


Figure 3.1: Discrete Markov chain. Reproduced from [Cal10].

the new combination of MCMC and importance sampling shown in Chapter 4.

The *state space*, $\mathcal{X}$, of a Markov chain may be discrete or continuous, however, in this work we will only treat the continuous case, and we think of $\mathcal{X}$ as the parameter space of a statistical model. Let X_t denote the random variable that is the chain output at time t. Specific instances of X, that is, elements of $\mathcal{X}$, are represented by x, y, z. For brevity, we do not use the vector notation $\boldsymbol{x}$, but $\mathcal{X}$ is certainly not restricted to 1D. The defining property of a Markov process is

$$P\left(X_t = x | X_0 = x_0, \dots, X_{t-1} = x_{t-1}\right) \equiv P\left(X_t = x | X_{t-1} = x_{t-1}\right) \ . \tag{3.5}$$

The one-step transition kernel P_{xy} to go from state x to state y,

$$P_{xy} = P(X_{t+1} = y | X_t = x) \ , \tag{3.6}$$

completely defines the Markov chain, along with the initial state X_0. Similarly, P^n_{xy} is the n-step transition kernel. For illustration, let us assume $\mathcal{X}$ has k elements, then the transition kernel is a matrix $\mathbf{P} \in \mathbb{R}^{k \times k}$, P_{xy} is the element in the x^{th} row and y^{th} column, the normalization condition reads $\sum_y P_{xy} = 1$, and the n-step transition kernel is simply the n^{th} matrix power $\mathbf{P}^n$. In fact, we can decompose each element of $\mathbf{P}^n$ as

$$P^n_{xy} = \sum_z P^r_{xz} P^{n-r}_{zy} \tag{3.7}$$

by introducing the intermediate state z, reached after $r \leq n$ steps. Equation (3.7), known as the Chapman-Kolmogorov equation, is the exact analogy of the path integral [Fey42] defined on a discrete spacetime. For consistency, we define $P^0_{xy} = \delta_{xy}$ with the Kronecker symbol δ, which naturally extends to the continuous case in which it denotes the Dirac distribution.

We shall need the following definitions. Two states *communicate* if there exists n such that there is nonzero probability to access one state from the other in n steps

$$P^n_{xy} \geq 0, P^n_{yx} \geq 0 \ . \tag{3.8}$$

Communication defines an equivalence relation on the state space; if all states in $\mathcal{X}$ are in the same equivalence class, the chain is *irreducible*. The *period* of a state, $d(x)$, is the greatest common divisor of all $n \geq 1$ for which the return probability P^n_{xx} is nonzero. If $d(x) = 1 \forall x$, the chain is called *aperiodic*. A chain is *stationary* if the joint distributions of

$$(X_t, X_{t+1}, \dots, X_{t+n}) \text{ and } (X_{t+h}, X_{t+1+h}, \dots, X_{t+n+h}) \tag{3.9}$$

agree for all h and arbitrary t. In the following, we will only consider stationary, aperiodic, and irreducible chains. Let f^n_{xx} denote the probability of a *first* return from x to x after exactly n steps. Then the *basic limit theorem* is

$$\lim_{n\to\infty} P^n_{xx} \equiv P(x) = \frac{1}{\sum_{t=0}^{\infty} t f^t_{xx}} , \tag{3.10}$$

$$\lim_{n\to\infty} P^n_{yx} = P(x) . \tag{3.11}$$

The basic limit theorem states that, asymptotically, the probability of being in a state x is independent of the starting point. In practice, it is of paramount importance to verify that this asymptotic regime is reached, since we can only ever perform a finite number of steps on a computer. A specific algorithm to perform this check is presented in Section 3.1.2. If we interpret t as time, then $\sum_{t=0}^{\infty} t f^t_{xx}$ is the average return time, thus the limiting probability $P(x)$ is given by the frequency $1/\sum_{t=0}^{\infty} t f^t_{xx}$. If $P(x) > 0$ for one x, then $P(y) > 0 \; \forall \; y \in \mathcal{X}$, and the chain is *strongly ergodic*. Again, the term "ergodic" is borrowed from physics; it describes the fundamental assumption of statistical mechanics that the time average of a macroscopic system equals the average over the microscopic states. From (3.7) and (3.10), we can immediately infer the left-eigenvalue problem

$$P(x) = \sum_z P(z) P_{zx} . \tag{3.12}$$

The set $\{P(x) : x \in \mathcal{X}\}$, is the *stationary* distribution of the Markov chain. The important message is this: with the help of the Metropolis-Hastings algorithm, we can construct the proper transition kernel such that the stationary distribution of the Markov chain is the function we want to sample from, the *target* density — the posterior in our applications. Thus the chain — the sequence of states visited — represents a (correlated) sample from the target provided the chain is ergodic. A sufficient condition for the distribution $P(x)$ to be the target distribution is *detailed balance* [RC04, Ch. 6]:

$$P_{xy} P(x) = P_{yx} P(y) . \tag{3.13}$$

Detailed balance reflects a state of equilibrium and symmetry; i.e., it is as probable to be in x and move to y as it is in the opposite direction to be in y and move to x.

The celebrated Metropolis-Hastings algorithm [Met+53; Has70] yields the right transition kernel to sample from a given target $P(x)$; at its core, it is deceptively short and simple. The key new ingredient is the *proposal* function $q(y|x)$, which generates a new proposal point y based on the current state x. The proposal is essential to the success of the method, and needs to be adjusted to the problem at hand. For now, our only requirement on q is that it have nonzero probability on the entire state space to guarantee irreducibility. Let us define the probability of accepting the proposal y as

$$\rho(y|x) = \min\left\{ \frac{P(y)}{P(x)} \cdot \frac{q(x|y)}{q(y|x)}, 1 \right\} . \tag{3.14}$$

Note that $\rho(y|x)$ is independent of the normalization of $P(x)$; it cancels in the factor $P(y)/P(x)$. Metropolis [Met+53] originally suggested to use a symmetric proposal, $q(x|y) = q(y|x)$, such that $\rho(y|x)$ is independent of q. The ability to use an asymmetric proposal was added later by Hastings [Has70], and therefore the factor $q(x|y)/q(y|x)$ is known as the Hastings factor. The Metropolis-Hastings algorithm determines the next point of the chain, X_{t+1}, as follows: generate $u \sim U_{[0,1]}$, then set

$$X_{t+1} = \begin{cases} y, & u < \rho(y|x) \\ x, & \text{else} . \end{cases} \tag{3.15}$$

Since proposals may be rejected, we may have repeated instances of x in the chain, thereby introducing autocorrelation, but also guaranteeing aperiodicity. We can decompose the transition kernel into two parts. Either we accept the new proposal point, or the chain reproduces its current state:

$$P_{xy} = \rho(y|x)q(y|x) + (1 - P_{x\cdot})\,\delta(y - x) \ , \tag{3.16}$$

where $P_{x\cdot} = \int \mathrm{d}z\, \rho(z|x)q(z|x)$ is the probability of jumping to an arbitrary state from x. We now want to verify that this kernel satisfies detailed balance (3.13). For $x = y$, this is trivial. In the nontrivial case, we have

$$\begin{aligned} P_{xy}P(x) &= \rho(y|x)q(y|x)P(x) = \min\left\{\frac{P(y)q(x|y)}{P(x)q(y|x)}, 1\right\} q(y|x)P(x) \\ &= \min\left\{P(y)q(x|y), q(y|x)P(x)\right\} \qquad (3.17) \\ P_{yx}P(y) &= \rho(x|y)q(x|y)P(y) = \min\left\{\frac{P(x)q(y|x)}{P(y)q(x|y)}, 1\right\} q(x|y)P(y) \\ &= \min\left\{P(x)q(y|x), q(x|y)P(y)\right\} \qquad (3.18) \\ \Rightarrow P_{xy}P(x) &= P_{yx}P(y) \ . \qquad (3.19) \end{aligned}$$

Thus by the basic limit theorem, $P(x)$ is the target distribution of the Markov chain that is reached asymptotically.

3.1.1 Adaptive Metropolis-Hastings

The basic properties of the Metropolis-Hastings algorithm guarantee a set of samples from the target only asymptotically. The overarching goal of Monte Carlo sampling is to construct an algorithm that can handle any target in a black box manner; i.e., no knowledge of the target is required initially, but the algorithm learns the details as it proceeds. Such an algorithm that works *efficiently* with any target, in any dimension, without the need of carefully checking consistency of the result etc. is the applied statistician's dream, but surely it does not exist. However, it serves as a guide to direct our efforts.

In this subsection, we describe the standard MCMC approach with a local, adaptive, multivariate proposal function based on [HST01; Wra+09]. For a *local* proposal $q(\boldsymbol{x}|\boldsymbol{y})$, the density at the proposed new point $\boldsymbol{x}$ depends on the current point $\boldsymbol{y}$, whereas a *global* proposal is independent of $\boldsymbol{y}$, $q(\boldsymbol{x}|\boldsymbol{y}) = q(\boldsymbol{x})$. The basic local random walk is performed with a symmetric proposal centered around y. It is a remarkable that in practice a simple class of proposal functions, such as the multivariate normal distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right) \ , \tag{3.20}$$

or the multivariate *Student's t* distribution

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma\left((\nu+d)/2\right)}{\Gamma\left(\nu/2\right)(\pi\nu)^{d/2}}\,|\boldsymbol{\Sigma}|^{-1/2}\left(1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-(\nu+p)/2}, \tag{3.21}$$

can adapt in such a way as to efficiently generate samples from essentially any smooth, unimodal distribution. The parameter ν, the degree of freedom, controls the "fatness" of the tails of $\mathcal{T}$; the covariance of $\mathcal{T}$ is related to the *scale matrix* $\boldsymbol{\Sigma}$ as $\frac{\nu}{\nu-2}\times\boldsymbol{\Sigma}$ for $\nu > 2$, while $\boldsymbol{\Sigma}$ is the covariance of $\mathcal{N}$. Hence for finite ν, $\mathcal{T}$ has fatter tails than $\mathcal{N}$, and for $\nu\to\infty$, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)\to\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Before delving into the details, let us clarify at least qualitatively what we mean by an *efficient* proposal. Our requirements are

1. that it allow to sample from the entire target support in finite time,
2. that it resolve small and large scale features of the target,
3. and that it lead to a Markov chain quickly reaching the asymptotic regime.

An important characteristic of Markov chains is the *acceptance rate* α, the ratio of accepted proposal points versus the total length of the chain. We argue that there exists an optimal α for a given target and proposal. If $\alpha = 0$, the chain is stuck and does not explore the state space at all. On the contrary, suppose $\alpha = 1$ and the target distribution is not globally uniform, then the chain explores only a tiny volume where the target distribution changes very little. So for some $\alpha\in(0,1)$, the chains explore $\mathcal{X}$ well.

How should the proposal function be adapted? After a chunk of N_{update} iterations, we change two things. First, in order to propose points according to the correlation present in the target density, the proposal scale matrix $\boldsymbol{\Sigma}$ is updated based on the sample covariance of the last n iterations. Second, $\boldsymbol{\Sigma}$ is multiplied with a scale factor c that governs the range of the proposal. c is tuned to force the acceptance rate to lie in a region of $0.15\le\alpha\le0.35$. The α range is based on empirical evidence and the following fact: for a multivariate normal proposal function, the optimal α for a normal target density is 0.234, and the optimal scale factor is $c = 2.38^2/d$ as the dimensionality d approaches ∞ and the chain is in the stationary regime [RGG97]. We fix the proposal after a certain number of adaptations, and then collect samples for the final inference step. However, if the Gaussian proposal function is adapted indefinitely, the Markov property (3.5) is lost, but the chain and the empirical averages of the integrals represented by (3.1) still converge under mild conditions [HST01].

The efficiency can be enhanced significantly with good initial guesses for c and $\boldsymbol{\Sigma}$. We use a subscript t to denote the status after t updates. It is often possible to extract an estimate of the target covariance by running a mode finder like MINUIT [Jam75] that yields the covariance matrix at the mode as a by product of optimization. In the case of a degenerate target density, MINUIT necessarily fails, as the gradient is not defined. In such cases, one can still provide an estimate as

$$\boldsymbol{\Sigma}^0 = \operatorname{diag}\left(\sigma_1^2,\sigma_2^2,\ldots,\sigma_d^2\right)\,, \tag{3.22}$$

where σ_i^2 is the prior variance of the i-th parameter. In the global fit detailed in Chapter 7, there are a lot of nuisance parameters whose posterior is similar to their prior. In that case, (3.22) is a very good starting guess. The updated value of $\boldsymbol{\Sigma}$ in step t is

$$\boldsymbol{\Sigma}^t = (1-a^t)\boldsymbol{\Sigma}^{t-1} + a^t\boldsymbol{S}^t\,, \tag{3.23}$$

where $\boldsymbol{S}^t$ is the sample covariance of the points in the t^{th} chunk and its element in the m^{th} row and n^{th} column is computed as

$$\left(\boldsymbol{S}^t\right)_{mn} = \frac{1}{N_{\text{update}} - 1} \sum_{i=(t-1)\cdot N_{\text{update}}}^{t\cdot N_{\text{update}}} \left((\boldsymbol{x}^i)_m - \widehat{E_P[(\boldsymbol{x})_m]}\right)\left((\boldsymbol{x}^i)_n - \widehat{E_P[(\boldsymbol{x})_n]}\right) . \tag{3.24}$$

The weight $a^t = 1/t^\lambda, \lambda \in [0,1]$ is chosen to make for a smooth transition from the initial guess to the eventual target covariance, the implied cooling is needed for the ergodicity of the chain if the proposal is not fixed at some point [HST01]. One uses a fixed value of λ, and the particular value has an effect on the efficiency, but the effect is generally not dramatic; in this work, we set $\lambda = 0.5$ [Wra+09].

We adjust the scale factor c as described in Algorithm 1. The introduction of a minimum and maximum scale factor is a safeguard against bugs in the implementation. The only example we can think of that would result in large scale factors is that of sampling from a uniform distribution over a very large volume. All proposed points would be in the volume, and accepted, so $\alpha \equiv 1$, irrespective of c. All other cases that we encountered where $c > c_{max}$ hinted at errors in the code that performs the update of the proposal.

Algorithm 1 Single update of the covariance scale factor. We use $\alpha_{min} = 0.15$, $\alpha_{max} = 0.35$, $\beta = 1.5$, $c_{min} = 10^{-4}$, and $c_{max} = 100$.

```
if $\alpha > \alpha_{max} \wedge c < c_{max}$ then
    $c \leftarrow \beta \cdot c$
else if $\alpha < \alpha_{min} \wedge c > c_{min}$ then
    $c \leftarrow c/\beta$
end if
```

3.1.2 Asserting convergence

When we discussed the basic limit theorem (3.10), the need for a method to assess when a chain has become stationary became apparent. There are two basic approaches: one can look at the autocorrelation properties of a single chain [RC04, Ch. 12], or one runs multiple chains from different starting points and declares convergence once these chains mix. For then, the chain output indeed is independent of the starting point. We select the multiple-chain approach, as it allows for trivial parallel execution on a multicore or multiprocessor computing architecture.

Our approach is based on Gelman and Rubin [GR92]; we review their approach, and indicate where we differ. The basic requirement of the chains' starting points is that they come from a distribution that is overdispersed with respect to the target density. Gelman and Rubin suggest to find the target modes $\boldsymbol{\mu}_j, j = 1, 2, \dots$ and the respective covariance matrix from the second derivative at $\boldsymbol{\mu}_j$, and to create a *mixture density* $\hat{P}(\boldsymbol{x})$ that roughly interpolates $P(\boldsymbol{x})$ with one Student's t component for each mode as

$$\hat{P}(\boldsymbol{x}) = \sum_j \alpha_j \mathcal{T}\left(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu\right) , \tag{3.25}$$

and $\nu \approx 4$ (thick tails). The starting points are drawn with *importance resampling* from $\hat{P}(\boldsymbol{x})$ with component weights w_j suitably adjusted. In contrast, we achieve a good compromise between overdispersion and efficiency by drawing $\boldsymbol{x}$ from the prior. In our main application, the global fit in Chapter 7, we assume a completely factorizable prior: the Wilson coefficients have a flat prior but the posterior is sharply concentrated, ensuring overdispersion. In other dimensions, the prior closely resembles the respective 1D marginalized posterior, thus the starting point is very close to a posterior mode.

The question that Gelman and Rubin pose is not: Have the chains converged? Instead, they seek to estimate how much information could be gained if the chains were run for an infinite number of steps. To this end, they consider a scalar quantity of interest, T, that depends on the parameters $\boldsymbol{x}$. Example quantities include one of the d parameters, $T = x_i$, or the logarithm of the target, $T = \log P(\boldsymbol{x})$.

Using a flat prior in T and treating the N samples from k chains as data, the posterior distribution $P_{kN}(T)$ denotes what is known about T. We stress that $P_{kN}(T)$ is the posterior for the inference problem described in the previous paragraph, and it is not to be confused with the target density from which samples are to be generated. $P_{kN}(T)$ includes uncertainty due to finite k and N, as well as due to the target distribution $P(T|\boldsymbol{x})$. It is assumed that, asymptotically, samples are from the target, $\lim_{N\to\infty} P_{kN}(T) \to P(T|\boldsymbol{x})$. One can incorporate the finite N effect in a Student's t distribution with scale $\hat{V}$ and degrees of freedom ν, and compare its variance to the asymptotic normal form (see Section 2.3) of the posterior with variance σ^2. Note that both distributions are estimated from the same set of $k \cdot N$ draws. Finally, the main quantity of interest is the Gelman-Rubin R *value*

$$R(T) = \frac{\hat{V}}{\sigma^2} \cdot \frac{\nu}{\nu - 2} \,. \tag{3.26}$$

The R value estimates the reduction of uncertainty about T which would result if $N \to \infty$. In theory, $R \lesssim 1.1$, and a value $R \approx 1$ indicates convergence. Note that R is a function of N, k, and the k individual chain means and variances of T. For a practical implementation, it is important to correctly deal with the special cases of $\hat{V} = 0, \sigma^2 \neq 0$ (R undefined), $\hat{V} = \sigma^2 = 0$ ($R = 1$), and $\nu < 2$ (R undefined) to avoid an infinite or NaN floating point value. In addition, since there is statistical uncertainty on $\hat{V}$ and σ^2, for small $N \lesssim 500$, R may actually come out slightly less than 1, say $\mathcal{O}\left(0.98\right)$.

3.1.3 Summary of adaptive MCMC

We now summarize the basic steps of our implementation of the adaptive MCMC in the software package EOS [Dyk+12]; see Algorithm 2. k chains are initialized with starting points drawn from the prior. The prior variances and a dimension-dependent scale factor are used to construct a multivariate normal distribution, serving as the initial proposal function for a local random walk. In each iteration, the proposal is centered around the current point. The chains are run for chunks of N_{update} iterations in parallel. The proposal covariance is adapted based on the N_{update} steps as in (3.23) and Algorithm 1. Convergence is declared, that is the *prerun* or *burn-in* is completed, if all chains have a suitable acceptance rate and the R value for each parameter is sufficiently close to 1. In addition, we require a minimum number of iterations, N_{min}, to avoid premature convergence, and conversely N_{max} to stop if convergence is not reached. For the *main run*, the proposal covariance is held constant, and a fixed number of samples $\mathcal{O}\left(10^5\right)$ is collected from each chain to perform inference on the parameters. The prerun samples are discarded.

Algorithm 2 The MCMC algorithm with prerun and main run. We use $N_{min} = 10^4$, $N_{max} = 5 \times 10^5$ and $N_{\text{final}} = 10^5$.

```
Require: number of chains $k$, $N_{\text{update}}$, proposal type = $\mathcal{N}, \mathcal{T}$
  for all chains do                                                    (▷) initialization
      Draw initial point from prior
      Setup proposal with diagonal initial scale matrix from priors
      Rescale matrix by $2.38^2/d$
  end for
  number of iterations $i \leftarrow 0$
  while $i < N_{max}$ do                                               (▷) prerun
      for all chains do
          run chains for $N_{\text{update}}$ iterations
          update proposal
      end for
      $i \leftarrow i + N_{\text{update}}$
      if acceptance rates and $R$ values OK and $i \geq N_{min}$ then
          prerun finished
      end if
  end while
  for all chains do                                                    (▷) main run
      run chains for $N_{\text{final}}$ iterations
  end for
```

3.2 Multimodal example

In order to illustrate the MCMC algorithm, we consider an explicit example target density P, similar in many aspects to the posteriors encountered in Chapter 7. P has multiple, well separated maxima, each of which has the same shape, but potentially a different probability mass. For the purpose of simplicity, we shall begin in $d = 2$ dimensions and assume that $P(\boldsymbol{x}) = P(x_1, x_2) = P(x_1) \cdot P(x_2)$. Let $P(x_1)$ a mixture of two LogGamma components (cf. Appendix A.2) centered around $x_1 = \pm 10$ with unit scale and unit shape parameter. Similarly, $P(x_2)$ is a mixture of standard normal distributions centered around $x_2 = \pm 10$, where we assume non-equal relative weight ω:

$$\begin{array}{rclcl} P(x_1) & = & \frac{1}{2}\text{LogGamma}(x_1|10,1,1) & + & \frac{1}{2}\text{LogGamma}(x_1|-10,1,1) \\ P(x_2) & = & \frac{\omega}{1+\omega}\mathcal{N}(x_2|10,1) & + & \frac{1}{1+\omega}\mathcal{N}(x_2|-10,1)\ . \end{array} \tag{3.27}$$

A single mode of $P(x_1, x_2)$ is depicted in Fig. 3.2, using the analytical values (Fig. 3.2(a)) and the MCMC histogram approximation (Fig. 3.2(b)) based on a single chain with 200 000 iterations, with the first 20 % discarded for burn-in, and proposal adaptation every 500 steps until the very end (no main run). The MCMC initialization is as follows. For concreteness, we assume flat priors on the ranges $(x_1, x_2) \in [-30, 20] \times [-20, 20]$, from which the initial point is drawn. The initial covariance of the Gaussian proposal in each dimension is that given by the respective prior (flat in this example), with an additional scale factor of $1/10$ on top of the dimensional scaling c that ensures initial acceptance rates of > 20 %.

The Markov chain in Fig. 3.2(b) quickly reaches the asymptotic regime and its stationary distribution is the part of the target $P(x_1, x_2)$ in the positive quadrant $x_1, x_2 > 0$. The

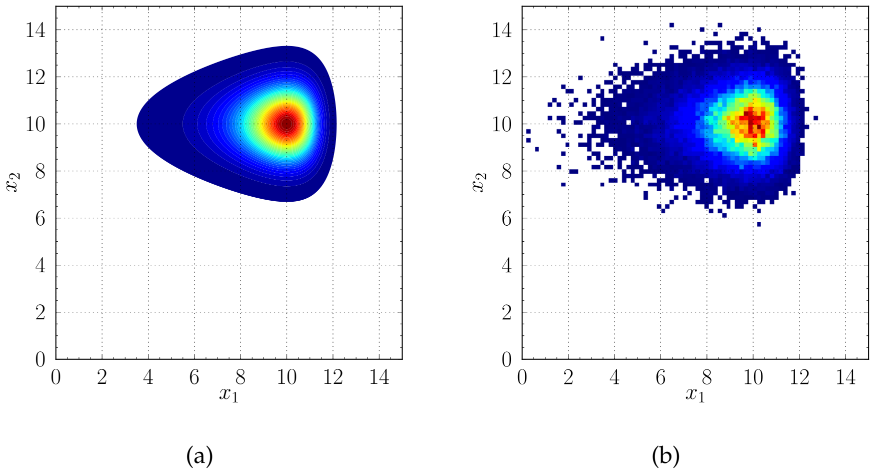


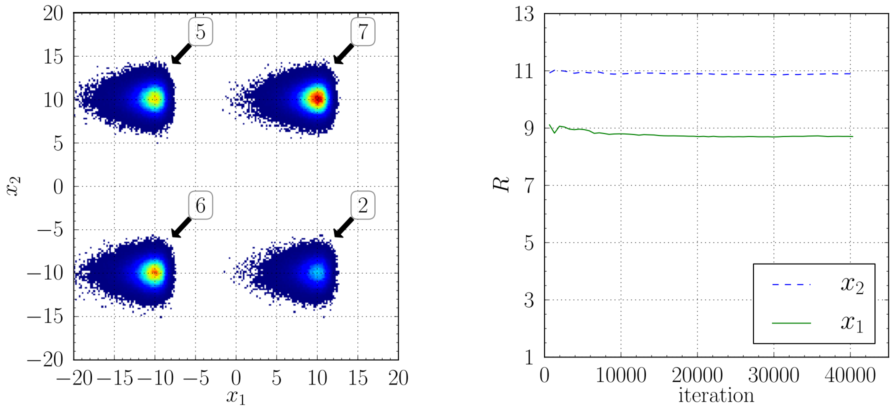
Figure 3.2: The example density $P(x_1, x_2) = P(x_1)P(x_2)$ with $P(x_1)$ and $P(x_2)$ given in (3.27). (a) The density contours from evaluating $P(x_1, x_2)$ on a fine grid around the $(+,+)$ mode. (b) Histogram of a single Markov chain with 200 000 iterations that visits only the $(+,+)$ mode.

chain does not visit the other three modes, thus the chain's stationary distribution is *not* the target; in this example, MCMC fails at global scale, but it is successful at the local scale.

What happens if we run $k = 20$ chains in parallel? The distance between the modes is deliberately chosen large in this example; each chain only visits a single mode. While the combination of chain outputs, shown in Fig. 3.3(a), gives a good impression of the local features of $P(x_1, x_2)$, it does not reflect the relative weight of the modes at all. Using fixed random number generators, we verified that the chains behave identically for drastically different values of ω; explicitly we checked this for $\omega = 1, 10, 10^4, 10^6$. The relative proportions of the maxima in Fig. 3.3 are entirely due to the number of chains in a particular mode, which in turn depends only on the chain's initial position and the random numbers used. The proportions do *not* depend on the value of $P(x_1, x_2)$ even though they should for correctness. We label the modes by their signature, for example the mode at $(10, 10)$ is the $(+,+)$ mode. For clockwise ordering $(+,+), (+,-), (-,-), (-,+)$, the relative proportions in the concrete run of Fig. 3.3(a) are given by 7:2:6:5.

The large R values, displayed in Fig. 3.3(b), for both x_1 and x_2 clearly show that the ensemble of chains has not mixed, therefore convergence cannot be declared. The difference between $R(x_1) \approx 9$ and $R(x_2) \approx 11$ is due to the smaller ratio of single-mode variance to intermode distance in x_1 direction, but the absolute value is not important; what matters is that $R(x_i) \gg 1$. In contrast, when considering the group of chains in one mode, both $R(x_1)$ and $R(x_2)$ are less than 1.1. We can thus conclude that the chains converge in each group, but the ensemble of chains fails in sampling from the target density. This is not a special case, MCMC methods are well known to fail when the support of the target is disconnected.

One may wonder if there is an error in the Metropolis-Hastings algorithm (cf. Section 3.1), its adaptive realization (cf. Section 3.1.1), or our implementation. The answer is no; the loophole is the condition of the "asymptotic" regime. There is nonzero prob-



(a)

(b)

Figure 3.3: The example density $P(x_1, x_2) = P(x_1)P(x_2)$ with $P(x_1)$ and $P(x_2)$ given in (3.27).
(a) MCMC results from combining 20 chains with 40 000 iterations each. The number of chains trapped in a mode is indicated.
(b) R values for x_1, x_2 as a function of the iteration stay well above $R = 1.1$, clearly indicating that the chains do not mix.

ability to propose and accept a new point in $(+, +)$ given a current point in $(-, -)$; the modes "communicate". But the tails of the Gaussian fall off quickly; a transition from one mode to the other is so unlikely that is does not occur in the finite-length chains that we recorded. Yet the transition occurs with probability one in an infinite number of iterations, and that is all that is needed for asymptotic correctness.

The task is then to construct a proposal that helps chains frequently jump from one mode to the other in order to correctly sample from the target distribution. In Appendix D, we present some attempts at a proposal q that "knows" about the modes and contains a nonlocal contribution to stimulate moves going beyond the reach of a typical local jump. However, these variants are successful only in $d \lesssim 10$, hence of no use for the global fit with $d = 30$. A powerful solution to correctly deal with the multimodality, based on importance sampling, is presented in Chapter 4. In particular, the proper results superseding Fig. 3.3(a) for the example target $P(x_1, x_2)$ are shown in Fig. 4.6(b).

3.3 Importance sampling

The starting point of importance sampling is to compute the fundamental expectation value (3.1)

$$E_P[f] = \int \mathrm{d}x P(x) f(x) \;, \tag{3.28}$$

where $f(x)$ is a generic scalar function of x. For simplicity, we consider only the 1D case $x \in \mathbb{R}$, the extension to $\boldsymbol{x} \in \mathbb{R}^d$ is straightforward. The simple, but extremely powerful idea is to turn the integral (3.28) into an expectation value under a proposal density q [NMU51; RC04] as

$$E_P[f] = \int \mathrm{d}x\, q(x) \frac{P(x)}{q(x)} f(x) = E_q\left[\frac{P}{q} f\right] . \tag{3.29}$$

The proposal q is required to be positive on the support of P to guarantee convergence. For N independent draws from the proposal, the estimator is

$$\widehat{E_P[f]} = \frac{1}{N}\sum_{i=1}^{N} w_i f(x^i), \qquad x \sim q\ , \tag{3.30}$$

with the *importance weight* $w \equiv P(x)/q(x)$. As with MCMC, it is (fortunately) possible to work with P unnormalized, then one has to replace w_i/N in (3.30) by the *self-normalized* important weight

$$\bar{w}_i = \frac{w_i}{\sum_{i=1}^{N} w_i}\ . \tag{3.31}$$

Note that (3.30) follows immediately from the fundamental Monte Carlo principle (3.2), and is identical if $q = P$, in which case $w_i \equiv 1$.

For practical reasons, q must be of a simple form to permit direct sampling. Note that, unlike in MCMC, the normalization of P can be inferred if q is properly normalized, because the value of $P(x^i)$ appears explicitly in (3.30) through the weight $w = P/q$. Suppose $P(x)$ is an unnormalized posterior distribution and let $f(x) = 1$, then the estimate of the normalization constant — the evidence Z (2.5) — is just the average unnormalized importance weight,

$$Z = E_q\left[\frac{P}{q}\right] \approx \frac{1}{N}\sum_i w_i \tag{3.32}$$

(3.29) is valid for any q whose support includes that of P, so which q should we choose? Suppose $q(x) = P(x)/Z$. The variance of the evidence estimator is [OZ00]

$$V_q[Z] = E_q\left[\frac{P^2}{q^2}\right] - \left(E_q\left[\frac{P}{q}\right]\right)^2 = \int \mathrm{d}x \frac{P(x)}{Z}\left[\frac{P(x)}{P(x)}Z\right]^2 - Z^2 = Z^2 - Z^2 = 0\ . \tag{3.33}$$

So if q is the normalized version of P, the estimator has minimum variance; but if we could sample from P directly, we would have no need to introduce q and to use importance sampling at all. However, the very important message is this:

$$\boxed{\text{Make } q \text{ as close as possible to } P \text{ for maximum efficiency.}} \tag{3.34}$$

For completeness, we have to remark that the optimal q depends on f. Robert and Casella [RC04, Ex. 3.8] present an example in which tail probabilities are estimated, and they show that a suitable proposal reduces the importance sampling estimator's variance by orders of magnitude compared to $q = P$. In this work however, we are interested in multiple f at the same time; e.g. all 1D and 2D marginal distributions of P and the evidence. In that case, $q \approx P$ is considered optimal.

In addition, it is important that the tails of q do not fall off more rapidly than those of P [Gew89]; i.e., P/q must be bounded from above, else outliers with very large w negatively affect the accuracy. In other words, we require that the estimator's variance be finite, which follows from

$$E_q\left[f^2\frac{P^2}{q^2}\right] < \infty\ . \tag{3.35}$$

3.3.1 Adaptive importance sampling

The great advantage of importance sampling over MCMC is that samples are drawn *independently* from q; thus the sampling, and in particular the evaluation of the posterior, can be massively parallelized, and there is no burn-in phase. At present, massive parallelization seems to be the only viable option to benefit from new computing infrastructures in order to increase the number of posterior evaluations per wallclock time, because the frequency of a single CPU has leveled off in the last 10 years at about 3 GHz and is not likely to increase in the future [Nor12].

The Metropolis-Hastings algorithm with an adaptive local random walk is robust in terms of choice of q, and it requires very little information to start. With no information present, MCMC adapts its multivariate Gaussian or Student's t proposal on the fly; cf. Section 3.1.1. In contrast, there is no such simple function that works well with an arbitrary target in importance sampling.

Adaptive importance sampling, or *population Monte Carlo* (PMC) [Cap+04; Cap+08; Wra+09; Kil+10], is a relatively new approach that tries to remedy the situation. The basic idea is to use a flexible proposal function, a mixture density of multivariate normal or Student's t distributions, and iteratively update its form to match the target density as closely as possible. In each iteration, not only one sample, but an entire *population* of samples is used, hence the name. That is, each iteration is a regular importance sampling step.

Typically, only the samples after the last proposal update are used for inference, so another kind of prerun emerges in which samples are discarded (cf. [Cor+12] for a formulation of adaptive importance sampling that combines samples from all steps). The option to massively parallelize the evaluation and the ability to naturally deal with degenerate and multimodal target distributions by scattering the mixture components apart make it a very powerful alternative to MCMC.

Let us now discuss the PMC algorithm in more detail. The proposal in step t is a mixture density

$$q^t(\boldsymbol{x}) = \sum_{j=1}^{K} \alpha_j^t q_j^t(\boldsymbol{x}|\boldsymbol{\xi}_j^t), \qquad q_j^t \in \{\mathcal{N}, \mathcal{T}\} \; . \tag{3.36}$$

q_j^t is a single multivariate component that depends on the simulation parameters collectively denoted by $\boldsymbol{\xi}_j^t$. For the Gaussian case, $q_j^t = \mathcal{N}$, we have $\boldsymbol{\xi}_j^t = \left(\boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t\right)$, and for the Student's t case $q_j^t = \mathcal{T}$, $\boldsymbol{\xi}_j^t = \left(\boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t, \nu\right)$. The set of (normalized) component weights is denoted by $\{\alpha_j^t\}, j = 1, \ldots, K$. Note that the component type, $\mathcal{N}$ or $\mathcal{T}$ (including ν), is fixed throughout a PMC run. The $\mathcal{T}$ components are preferred if the target has degeneracies or fat tails.

It is important to quantitatively assess the distance between the target and the proposal in the spirit of (3.34) to have a well posed optimization problem; this is accomplished with the Kullback-Leibler divergence, or relative entropy, KL, an extension of the Shannon entropy (2.24) to continuous distributions [KL51]:

$$\mathrm{KL}\left(P \| q\right) = \int \mathrm{d}x P(x) \log \frac{P(x)}{q(x)} \; . \tag{3.37}$$

Beware that in general $\mathrm{KL}\left(P \| q\right) \neq \mathrm{KL}\left(q \| P\right)$, so KL is not an actual metric on the space of probability distributions. The minimum occurs for $P = q$, then $\mathrm{KL}(P \| P) = 0$. KL is an important information-theoretical quantity; it quantifies the expected number of extra bits needed to code up a message — random sample — from P using an alphabet with characters distributed according to q, rather than using the true distribution P.

The goal in each update step is to reduce $\mathrm{KL}\,(P\|q)$. The general problem of optimizing the KL functional is intractable, it is therefore necessary to reduce the complexity to an ordinary parameter optimization problem by fixing q to the form (3.36) and optimizing over $\alpha_j, \boldsymbol{\xi}_j, j = 1, \ldots, K$. We want to remark that in the basic formulation of [Cap+08], the parameter ν of the Student's t distribution is held fixed, but it could be updated along with the other parameters through 1D numerical optimization [HOVD11].

KL is minimized using a variant of the expectation-maximization algorithm; see [Cap+08; Wra+09] for details, and Appendix E for a general introduction to expectation maximization. In the limit that the number of samples per update step tends to ∞, the algorithm guarantees that $\mathrm{KL}(P\|q^{t+1}) < \mathrm{KL}(P\|q^t)$. For the Gaussian and Student's t case, the updated values α_j^{t+1} and $\boldsymbol{\xi}_j^{t+1}$ are known, relatively simple-to-evaluate expressions of the importance sampling output $(x_i^t, w_i^t) : i = 1 \ldots N$ and q^t [Cap+08]. But they require at least a single summation over all samples, thus if N is large, parallelization of the proposal update is beneficial.

The update procedure is nothing but fitting q to P via the set of importance samples $\{(x_i^t, w_i^t) : i = 1 \ldots N\}$. Let us consider an example in $d = 30$ dimensions with $K = 100$ components. Then q contains a total of

$$\dim \alpha^t + \dim \boldsymbol{\xi}^t = 100 + (30 + 30^2) \times 100 = 93\,100 \tag{3.38}$$

parameters. It is thus necessary to determine an appropriate value for N; on the one hand, it should be large to have as much information about the target as possible, and certainly $N \gtrsim \dim \boldsymbol{\xi}^t$, on the other hand, it should remain small as the CPU time grows linearly with N. Fortunately, the wall clock time grows only as N/#processors in a parallel computing environment.

It is important to stress that PMC depends crucially on the initial proposal q^0, because the updates tend toward the next local minimum of KL, and not every such minimum leads to good sampling results. In addition, finding a good initial proposal is far from trivial; we assume it given for the time being, but we devote Chapter 4 to explain a new algorithm that forms q^0 automatically. In our opinion, the absence of such an initialization procedure until now made PMC basically unusable with very complicated target densities.

3.3.2 Convergence monitoring

Is there a quantity similar to the R value to use as a stopping criterion? Two statistics are available. Obviously, if $\mathrm{KL}(P\|q) = 0$, or equivalently if $\exp\left(-\mathrm{KL}(P\|q)\right) = 1$, no further improvement can be made. The latter is estimated by the *normalized perplexity*

$$\mathcal{P} \equiv \exp\left(H^t\right)/N \ , \tag{3.39}$$

with the Shannon entropy (2.24) for the normalized weights in step t given by

$$H^t(\bar{w}_1^t, \ldots, \bar{w}_n^t) = -\sum_i \bar{w}_i^t \log \bar{w}_i^t \ . \tag{3.40}$$

By construction, $\mathcal{P} \in [0,1]$. The other relevant quantity is the *effective sample size* (ESS), heuristically motivated in [LC95]. While $\mathcal{P}$ is sensitive rather to the mean of the distribution of the importance weights, ESS is a function of variance. Define the *coefficient of variation* as

$$C^2 = \frac{1}{N} \sum_{i=1}^{N} \left(N \bar{w}_j^t - 1 \right)^2 , \tag{3.41}$$

and note that for large N, C^2 is a reasonable approximation to $V[\bar{w}^t]$. Suppose that N_0 of the weights vanish, while the other $N - N_0$ weights are equal and nonzero. Then $C^2 = N/(N - N_0) - 1$, and the normalized ESS is

$$\text{ESS} \equiv \frac{1}{1 + C^2} = \frac{N - N_0}{N} . \tag{3.42}$$

Thus ESS is an estimate of the fraction of samples that effectively contribute to the importance sample, ESS $\in [0,1]$. Ideally, all weights are identical, then ESS $\equiv 1$.

An outlier — a sample with a weight much larger than the average — has a particularly large impact on ESS, whereas $\mathcal{P}$ is more robust against outliers. We will see in Section 4.3.3 that individual outliers are clearly visible in density plots of marginal distributions, and the ESS provides a quantitative warning about existing outliers. On the contrary, contours inferred from the marginal distributions are affected to a lesser extent by outliers. Thus, if the contours are the primary interest, the important criterion is $\mathcal{P}$. Hence, we do not to use ESS for the stoppage rule. A more detailed reasoning for this is given below in Section 4.3.3. The PMC algorithm in abstract form, including our stoppage rule, is given in Algorithm 3.

Algorithm 3 The generic PMC algorithm. We use $t_{min} = 2$, $t_{max} = 20$, $\varepsilon = 0.02$, $\mathcal{P}_{crit} = 0.92$, and $N_{\text{final}} = 2 \times 10^6$

```
Require: number of samples $N$, initial proposal $q^0$
  converged ← false
  while $(t < t_{max}) \wedge (\neg\text{converged})$ do                 ( ▷) Update loop
    $\{(\boldsymbol{x}_i, w_i)\}_{i=1}^{N}$ = IMPORTANCE_SAMPLE($q^t, N$)
    if $t > t_{min} \wedge \left( \left| \frac{\mathcal{P}^t - \mathcal{P}^{t-1}}{\mathcal{P}^t} \right| < \varepsilon \vee \mathcal{P}^t > \mathcal{P}_{crit} \right)$ then
      converged ← true
    end if
    $q^{t+1}$ = UPDATE_PROPOSAL($q^t, \{\boldsymbol{x}_i, w_i\}_{i=1}^{N}$ )
    $t \leftarrow t + 1$
  end while
  if converged then                                                        ( ▷) final step
    $\{(\boldsymbol{x}_i, w_i)\}_{i=1}^{N_{\text{final}}}$ = IMPORTANCE_SAMPLE($q^t, N_{\text{final}}$)
  end if
```

4 Markov Chains and adaptive importance sampling united

The simple example of the multimodal target density introduced in Section 3.2 clearly shows the weaknesses of the local random walk MCMC approach. If the various modes are well separated, a single chain visits only a single mode. In particular, which mode is visited does not depend on the probability mass of the mode; instead, the chain's initial position and the random numbers used in each Metropolis-Hastings step are the decisive factors.

On the contrary, the example also shows the strengths of MCMC. Given a sufficiently large number of chains, k, there is good chance that every mode is covered by at least one chain. Running several chains in parallel is not really a burden, as it is the simplest and most natural means to parallelize the evaluation of the target density, and it does not increase the required wallclock time. Without extra information required from the user, we can adapt a simple multivariate proposal function to draw samples from the target, at least in a subset of the full parameter space. Thus every chain provides valuable local information, and the crucial task is to combine the information of the ensemble of chains into a global picture.

Parallelization of MCMC is limited due to the sequential nature of the Metropolis-Hastings algorithm. Given the number of iterations required for burn-in and adaptation of the proposal, a minimum of $\mathcal{O}\left(10^4\right)$ steps are required. Hence if we have 50 cores available, we can run $k = 50$ chains, but the wallclock time is that needed for a single chain on a single core for at least $\mathcal{O}\left(10^4\right)$ iterations.

The population Monte Carlo (PMC) algorithm offers the complementary set of strengths and weaknesses compared with MCMC. First, it features the ability to massively parallelize the posterior evaluation to reduce the wallclock time by using more cores. Second, its mixture proposal function naturally copes with multiple modes. Third, importance sampling yields posterior samples along with the evidence. However, the main disadvantage of PMC is the crucial dependence on the initial proposal. In the first paper about PMC in physics [Wra+09], the authors discuss several rather basic strategies to place the initial components, such as centering in the allowed range or around a mode, with covariance taken from the inverse Hessian at the mode, For the complicated posteriors that we treat in the global fit, for an example see Fig. 7.1, these strategies failed completely.

Our focus is on creating a good initial proposal for PMC with a minimum of manual intervention. Our new suggestion is to *combine the best of MCMC and PMC in three steps* (cf. Fig. 4.1):

1. We perform a prerun with k Markov chains in parallel, each doing a local random walk with an adaptive proposal. Then we extract the support of the target density by splitting each chain into many small patches.

2. Sample mean and covariance of each patch define one multivariate density, the

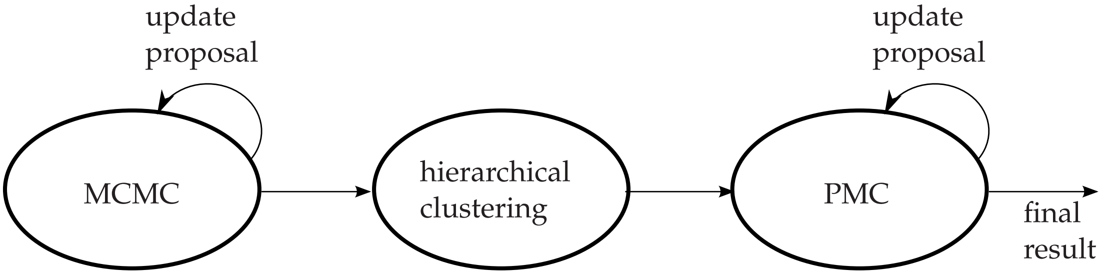


Figure 4.1: Schematic flow of information to construct a good initial proposal for PMC based on a MCMC prerun and hierarchical clustering.

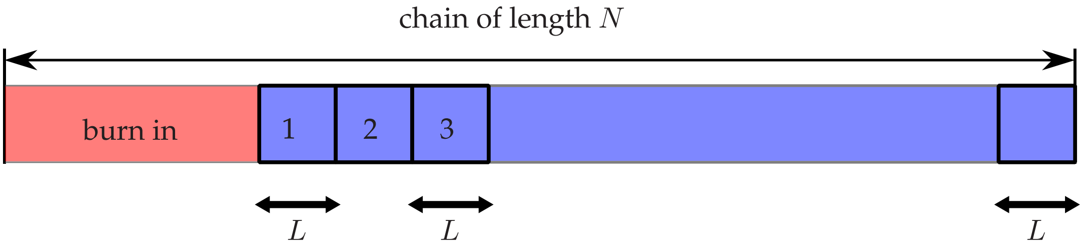


Figure 4.2: Partitioning of a chain into patches of length L to extract local information about the target from the chain's random walk.

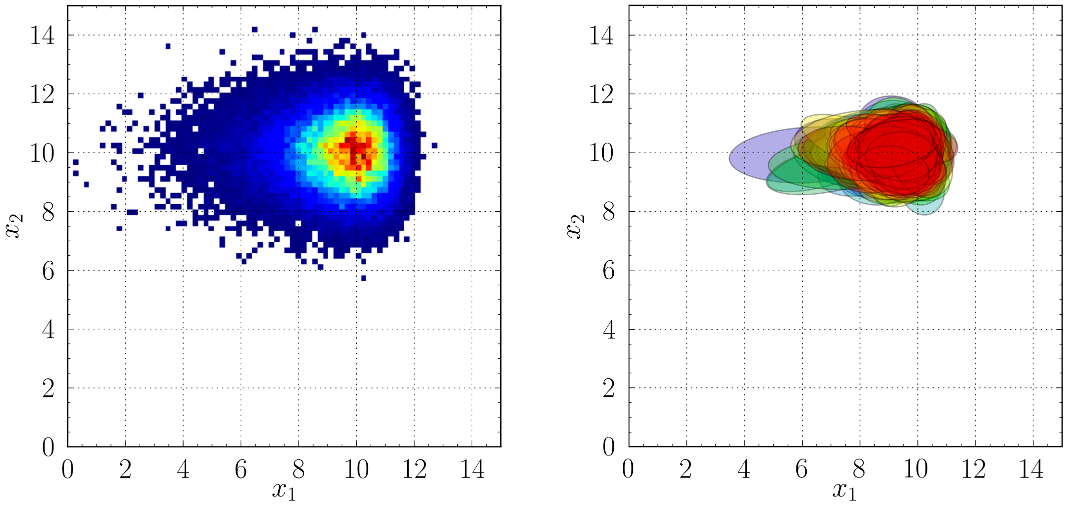
collection of patches yields a mixture density. Typically, there are more patches than actually needed; hierarchical clustering produces a mixture with far fewer components but essentially the same knowledge of the target density by removing redundant information.

3. The output of hierarchical clustering is used with minor modifications as the initial proposal for PMC. We then run the standard PMC updates until convergence.

The combination of MCMC and PMC is one leap forward towards a black-box Monte Carlo sampler that learns the relevant features of the target density automatically. In order to make optimal use of a parallel computing infrastructure, the MCMC prerun is kept to a minimum length, and most evaluations of the target density are performed at the PMC stage. In the remaining sections, we describe the three stages in detail, and give ample guidance to deal with the practical issues that arise along the way. First, we explain how we run the Markov chains. Second, we present hierarchical clustering, the link that connects MCMC and PMC. Next, we discuss the effects of various settings on the PMC run. Then we summarize how to tune algorithm parameters, and finally present an outlook of possible improvements. All along, we illustrate the algorithm by evolving the multimodal example from Section 3.2.

4.1 Markov chain prerun

Assuming no knowledge of the target density other than that is zero outside of a given hyperrectangle in $\mathbb{R}^d$, we draw the initial positions of the Markov chains from the uniform distribution. If the target is a posterior density and the priors are of a simple



(a)

(b)

Figure 4.3: (a) Histogram of a single Markov chain with 200 000 iterations that visits only the $(+,+)$ mode (Fig. 3.2(b)). (b) Centers and marginal 1-σ contours of multivariate Gaussian components computed from short chain patches.

form, we draw the starting points directly from the prior. In most problems, the prior is significantly more diffuse than the posterior, hence this choice of seeding the chains automatically satisfies the overdispersion condition of the R value (cf. Section 3.1.2). But the main reason why overdispersion is desirable to us is not the validity of the R value distribution. Instead, in the face of potentially several maxima, with little analytical knowledge of a posterior that is available only in the form of computer code, it is imperative to explore the full parameter space and to find *all* regions of significant probability. These regions are not limited to local maxima, but include degenerate regions as well. Therefore, the number of chains, k, should be chosen significantly larger than the number of expected maxima. In the example problem with four maxima, we choose $k = 20$, and in the global fit $k = 40 - 60$.

If the modes are well separated, the R value remains far above values of $R = 1.1$, and convergence is not declared. Therefore, a suitable criterion when to terminate the prerun has to be chosen. In this work, we select a maximum number of iterations, N. Assuming the chains are efficient from the beginning, we choose $N = 40\,000$ for the example with $d = 2$, and discard the initial 20 % for burn-in. More iterations are needed in higher dimensions or with a target showing significant degeneracies, such that one chain needs longer to cover a single region of high probability. In our work, the largest value of N we use is 100 000 in $d = 31$ in the global fit with wide priors; cf. Section 7.2.1.

Given the prerun of k chains, we now extract the local information by exploiting the slow, diffusion-like exploration of the Markov chain. To this end, we choose a *patch length* L, and partition the history of each chain, modulo the burn-in, into patches of length L; see Fig. 4.2. For the i^{th} patch, we compute the sample mean $\boldsymbol{\mu}_i$ and sample covariance $\boldsymbol{\Sigma}_i$, and form a multivariate Gaussian density (see Fig. 4.3)

$$f_i(\cdot\,|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \mathcal{N}(\cdot\,|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\ . \tag{4.1}$$

The patch length ought to be chosen in such a way that small-scale features of the posterior can be explored during L iterations. Again, a good value of L increases with d and possible degeneracies. On the other hand, L must not be too small, else the chain cannot move enough. Combing all k chains together, we obtain a Gaussian mixture density of $\tilde{K}$ components

$$f(\cdot) \equiv \sum_{i=1}^{\tilde{K}} \alpha_i \mathcal{N}(\cdot\,|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{i=1}^{\tilde{K}} \alpha_i f_i(\cdot) \ , \tag{4.2}$$

and assign equal weight $\alpha_i = 1/\tilde{K}$ to each component. Note that we do *not* take into account the value of the posterior in each patch; rather, we will ultimately rely on PMC to find the proper component weights. In this way, an identical number of samples is drawn from each component in the initial, and most important, PMC step. For the example in $d = 2$ with $L = 100$ and the prerun settings described in Table 4.1, Row I, we have the enormous number of $\tilde{K} = 20 \times 320 = 6400$ components. This number is impractically large even with massive parallelization. Fortunately, we do not actually need all these components, as there is a lot of redundant information; two chains that mix will approximately yield the same information as a single chain. Further, if a single chain explores the same region multiple times during N iterations, then that chain itself is redundant.

4.2 Hierarchical clustering

It is important to reduce the complexity of the proposal mixture density to keep the number of samples needed in each PMC update step low for computational efficiency. At the same time, we wish to preserve as much information as possible, in particular where the support of the target density is. In the previous section, the mixture density f, defined in (4.2), contained a total of $\tilde{K}$ components, each resulting from one patch of a Markov chain. For the simple 2D, multimodal example, $\tilde{K} = 6400$ components are much more than needed for PMC. Our goal in this section is to compress the $\tilde{K}$ components into a mixture with only $K \ll \tilde{K}$ components by removing redundant information. *Hierarchical clustering* [GR04] is our weapon of choice. Compression is to be understood in the information-theoretical sense, because a distance measure between the mixtures based on the Kullback-Leibler divergence (3.37) is minimized during the clustering.

4.2.1 Review of hierarchical clustering

We start with a mixture density of $\tilde{K}$ Gaussian input components in d dimensions (4.2). The objective is to reduce the number of input components to $K < \tilde{K}$, resulting in a Gaussian mixture density

$$g(\cdot) \equiv \sum_{j=1}^{K} \beta_j g_j(\cdot) \ . \tag{4.3}$$

This goal is achieved by finding the optimal output component weights, means, and covariances of a K-component Gaussian mixture density minimizing the distance measure

$$d(f, g) = \sum_{i=1}^{\tilde{K}} \alpha_i \min_j \mathrm{KL}(f_i \| g_j) \ . \tag{4.4}$$

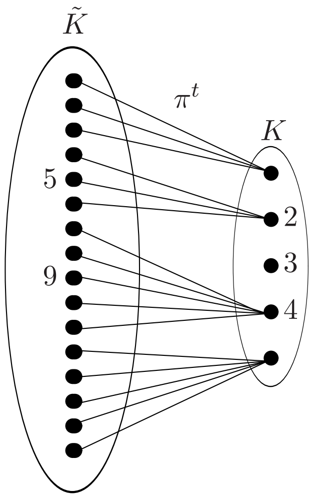


Figure 4.4: Hierarchical clustering. π^t maps each of the $\tilde{K}$ input components to the respective output component in step t with the smallest Kullback-Leibler divergence. For example, $\pi^t(5) = 2$ and $\pi^t(9) = 4$. The third output component is dead.

Due to the properties of the Kullback-Leibler divergence KL, $d(f,g) \neq d(g,f)$, hence $d(\cdot,\cdot)$ is not a distance in the strict sense. With our motivation of removing redundant information, constructing $d(\cdot,\cdot)$ based on KL is a natural choice, as KL quantifies the extra information needed to represent f_i by g_j; cf. (3.37). The particular form of $d(\cdot,\cdot)$ is chosen to facilitate the minimization of d. However, it can be shown that $d(\cdot,\cdot)$ arises as the limiting form of a likelihood of a suitable statistical model [GR04, Sec. 4]. Another very desirable property of $d(\cdot,\cdot)$ is that it only requires the exactly known Kullback-Leibler divergence between two Gaussians, whereas KL of a Gaussian mixture is not known analytically. Thus we do not need to work at the level of samples; once the components are defined from the patches, we can operate at the component level. This provides a dramatic speed up compared to similar hierarchical grouping procedures that require resampling data points. For two Gaussians, we have

$$\mathrm{KL}(1\|2) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \mathrm{Tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T \boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) - d\right]. \tag{4.5}$$

There is no analytical solution to find the optimal g, but an iterative algorithm, based on the expectation-maximization (EM) algorithm [DLR77], exists. It is a direct extension of the Gaussian mixture EM from the sample to the component level; the former is described in detail in Appendix E. In hierarchical clustering, the key in each step t is to find a mapping π^t to associate each of the $\tilde{K}$ input components with the closest output component, $\pi^t(i) \equiv \arg\min_j \mathrm{KL}(f_i \| g_j^t)$, as displayed schematically in Fig. 4.4.

Given the mapping π^t and its inverse $\pi^{-1,t}$, the parameters of the mixture g^t are now updated such that for each j with $\pi^{-1,t}(j) \neq \varnothing$, the new component g_j^{t+1} is the weighted

Algorithm 4 The hierarchical clustering algorithm. We use $\varepsilon_{min} = 10^{-4}$.

```
  $t \leftarrow 0$
Require: initial output components $\{g_j^0\}$
  repeat
      create optimal mapping $\pi^t$                                  ( ▷) regroup, E step
      compute $d^t(\pi^t, \{f_i\}, \{g_j^t\})$
      $t \leftarrow t + 1$
      for each output component do
          update weight, mean, and covariance                        ( ▷) refit, M step
      end for
  until $t > 1 \wedge \left(| (d^{t-1} - d^{t-2}) / d^{t-1} | < \varepsilon_{min}\right)$
```

average of all input components $f_i \in \pi^{-1,t}(j)$ with parameters

$$\beta_j^{t+1} = \sum_{i \in \pi^{-1,t}(j)} \alpha_i \ , \tag{4.6}$$

$$\boldsymbol{\mu}_j^{t+1} = \frac{1}{\beta_j^{t+1}} \sum_{i \in \pi^{-1,t}(j)} \alpha_i \boldsymbol{\mu}_i \ , \tag{4.7}$$

$$\boldsymbol{\Sigma}_j^{t+1} = \frac{1}{\beta_j^{t+1}} \sum_{i \in \pi^{-1,t}(j)} \alpha_i \left(\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^{t+1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^{t+1})^T\right) \ . \tag{4.8}$$

Note that both π^t and $\{g_j^t\}$ are updated, one after the other, in each step t, but the input components $\{f_i\}$ are fixed. Once $\pi^{-1,t}(j) = \varnothing$, the j-th output component will not be updated anymore in subsequent steps. Its weight is zero, hence it is "dead". Next, the optimal mapping π^{t+1} for the output components $\{g_j^{t+1}\}$ is constructed, the output components are updated again, and so forth. The process of *regrouping* and *refitting* is repeated until a minimum is found. Due to the discrete nature of the problem, a local minimum of $d(\cdot,\cdot)$ is reached after a *finite* number of steps when $\pi^{t+1} = \pi^t \Rightarrow d^{t+1} = d^t$. Observing that $d(\cdot,\cdot)$ changes only little near the minimum, it is even quicker to check the relative precision, $\varepsilon = | (d^{t+1} - d^t) / d^{t+1} |$, and to declare convergence if $\varepsilon \leq \varepsilon_{min} = \mathcal{O}\left(10^{-4}\right)$. In practice, the procedure, summarized in Algorithm 4, usually terminates due to $\pi^{t+1} = \pi^t$ for K small, and due to a small relative change of d^t for $K \gtrsim 20$.

4.2.2 Initialization

Hierarchical clustering, being an expectation-maximization variant, converges only on a local minimum of the distance measure $d(\cdot,\cdot)$. Given a large number of input components, there exist numerous local minima, hence it is crucial to supply good initial guesses for the output components $\{g_j^0\}$, such that the initial solution d^0 is already very close to a *good* final solution. We then note rapid convergence after $\mathcal{O}(10)$ steps. There are two important questions we need to address.

1. Where to put the initial output components $\{g_j^0\}$?
2. How many output components, K, are needed?

Unfortunately, we cannot offer a procedure to automatically calculate K. Goldberger and Roweis [GR04] vaguely recommend to use "standard methods for model selection". We can only speculate that they refer to the Bayesian information criterion

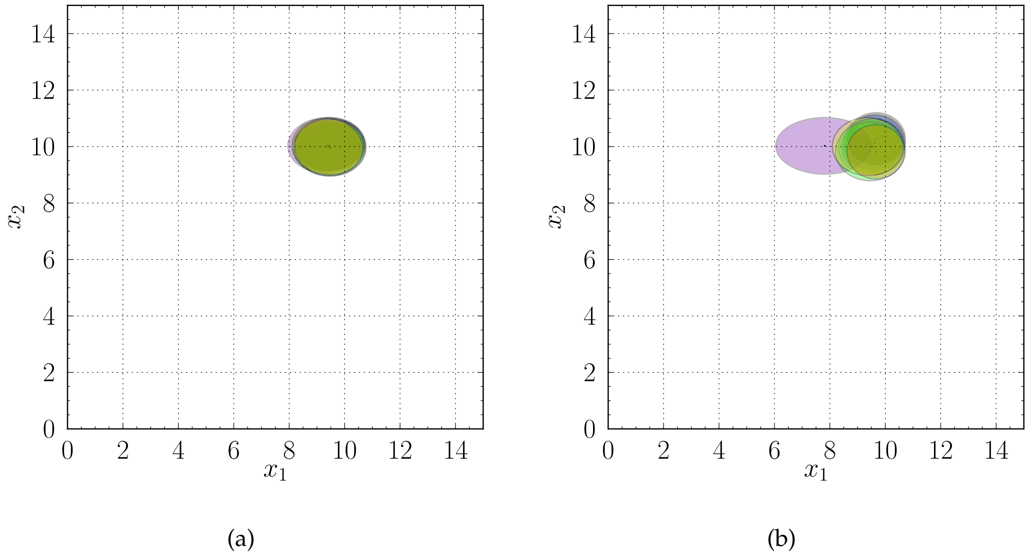


Figure 4.5: Centers and marginal 1-σ contours of an 8-component mixture density in the (a) initial guess and (b) final step of hierarchical clustering. We display only the $(+,+)$ mode of the multimodal example from Section 3.2. The tail of $P(x_1)$ is captured by the purple component; cf. Fig. 3.2(a).

[Sch78] or the Akaike information criterion [Aka74]. Another approach would be to add one component at a time until K is "large enough". It then remains to specify a quantitative stopping criterion. In [HOVD11], an attempt at such a criterion is presented, but it appears highly inefficient for large d and well-separated modes. More alternatives are discussed in Section 4.5.

With no algorithmic solution at the moment, we assume the user provides a sensible value of K. In Section 4.3, we will explore the effect of varying K on the PMC run to guide the user in this choice; as a rule of thumb, we recommend K should be at least as large as d.

But to answer the first question, we have a good idea *where* to place the components. The key is to group the chains, and to have a fixed number of components per group from *long patches*. To begin with, it is necessary to determine which chains have mixed in the prerun. Two or more chains whose common R values are less than a given constant, say $R_c = 1.1$, for all parameters, form a *group* of chains. Most importantly, this ensures that a similar and sufficient number of components is placed in every mode of the target density, regardless of how many chains visited that mode. We ignore the burn-in samples of each chain as described above in Section 4.1.

Let us assume we want K_g components from a group of k_g chains. If $K_g \geq k_g$, we find the minimal lexicographic *integer partition* (see [BC11] and references therein) of K_g into exactly k_g parts. Hence, the partition, represented as a k_g-dimensional vector of integers $\boldsymbol{n}$, is given by

$$\boldsymbol{n} = \left(\left\lceil \frac{K_g}{k_g} \right\rceil, \ldots, \left\lceil \frac{K_g}{k_g} \right\rceil, \left\lfloor \frac{K_g}{k_g} \right\rfloor, \ldots, \left\lfloor \frac{K_g}{k_g} \right\rfloor \right) , \tag{4.9}$$

where we used the ceiling ($\lceil\rceil$) and floor ($\lfloor\rfloor$) operation. The first $K_g \mod k_g$ parts are one larger than the remaining parts. For example, with $K_g = 6$ and $k_g = 4$, the partition is $(2, 2, 1, 1)$. If $K_g < k_g$, the integer partitioning cannot be performed as above. Instead, we combine all individual chains into one long chain, and set $k_g = 1$.

Finally, the i^{th} chain is partitioned into n_i long patches, and the sample mean and covariance of each patch define one multivariate Gaussian as before. The long patches, say there are two or three per chain, represent expectation values over a long time. Small-scale features are averaged out, while the center of gravity is preserved. Thus the initial output components from one group are very similar (cf. Fig. 4.5(a)), and the hierarchical clustering shifts and shrinks them to fit (cf. Fig. 4.5(b)). Due to the initial similarity, very few, and usually zero, components "die" during the hierarchical clustering. Thus, the chosen value of K is preserved, which is desired behavior.

In conclusion, let n_g denote the number of chain groups, then the initial mixture of output components for hierarchical clustering consists of $K = (K_g \times n_g)$ components. Note that n_g is determined automatically, but n_g is a function of the critical R value R_c, a parameter that requires moderate tuning. For well separated modes, $R_c = 1.1 - 2$ gives stable, reproducible results. As a rule of thumb, R_c can be chosen smaller when either the parameter space is low dimensional, or the distributions are narrow in the sense that the chains had sufficient time to fully explore one mode in N steps. In our 2D example with $N = 40\,000$ iterations in the prerun, $R = 1.1$ is fine, but if we extend the example by adding, for instance, 28 unimodal Gaussians directions, $R = 1.5$ is more suitable.

4.3 Population Monte Carlo

The result of the first two stages of the new algorithm, the MCMC prerun in Section 4.1 and the hierarchical clustering in Section 4.2, is a Gaussian mixture density g. Naïvely, we would set the initial proposal $q^0 = g$, and start mapping the target density with PMC. However, a number of considerations have to be taken into account. We use Gaussians because the hierarchical clustering is then particularly fast and simple to implement. But we do not expect the chain patches turned into Gaussians to approximate the target density with the highest precision. In particular, most realistic problems have thicker tails, and are more accurately described by a Student's t mixture[1]. In fact, a more complicated hierarchical clustering for Student's t exists [EAPG09], but we don't expect it to reduce the number of PMC updates. The sole purpose of g is to cover the support of the target with some accuracy, and the actual adaptation is left to the PMC update algorithm. In the end, we only use the samples drawn from the adapted PMC proposal for inference. We therefore consider it appropriate to perform two modifications to g.

First, all component weights are set equal, to balance the effect of an unequal number of chains in each group. The weights are adjusted properly in the first PMC update, so components are discarded if their target probability mass is low, and not because few chains visited them.

Second, if a Student's t mixture is believed to yield a better representation of the target, we create a "clone" of g where each Gaussian component is replaced by a Student's t component with identical location and scale parameter. The degree of freedom, ν, is

[1] See [DE10] for a review of the 2008 financial crisis with regard to the failure of Gaussian-based risk models to capture extreme events.

Algorithm 5 Initialization of the output components for hierarchical clustering. We use $a = 0.2$ and $R_c = 1.1 \ldots 2$.

```
Require: Empty initial mixture density $g^0$
Require: $k$ chains with $N$ samples
Require: Number of components per group $K_g$
  Discard the first $a \times N$ burn-in samples
  GROUP_CHAINS($R_c$)
  $n_g \leftarrow$ number of groups
  for all groups do
      $k_g \leftarrow$ number of chains in group
      if $K_g < k_g$ then
          Merge chains into one long chain
          $k_g \leftarrow 1$
      end if
      $\boldsymbol{n}$ =MINIMAL_PARTITION($K_g, k_g$)
      for all chains in group do
          Partition into $n_i$ patches
          for all patches in chain do
              Compute sample mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
              Add one component $g_j^0(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to mixture
          end for
      end for
  end for
  Assign equal weight $\frac{1}{n_g K_g}$ to every component
```

the same for all components, and currently has to be chosen a-priori by the user in the PMC approach. Its optimal value in the update is not known in closed form [Cap+08]. However, as noted in [HOVD11], ν can be obtained from one-dimensional root finding. This is one source of future improvement, as guessing the proper value of ν is not easy. In low dimensions, the difference is usually small, but for large d, the impact may be significant (see Section 4.3.2).

Assuming that q^0, the initial proposal, is fixed, there is still an open question before we can start PMC: how many samples N to draw from the proposal? In the derivation of the PMC update step, $N \to \infty$ is assumed, and this guarantees a reduced Kullback-Leibler divergence [Cap+08]. Large N ensures many samples from each component, but increases the computational burden. If N is too small, the updates may render q^{t+1} worse than q^t, and the PMC algorithm fails. A proper choice of N depends mostly on the dimensionality of the target density d; for guidance, cf. the discussion in Section 4.4. After all, N is a required input to PMC, and cannot be deduced from the problem. A reliable, quantitative rule to determine N would be very desirable, but is not available to us. We then attempt to ensure that every component is explored initially, so the quantity of interest is N_c, the *number of samples per component* in the first step, whence $N = K \times N_c$. Once the component weights are adjusted in the first update step, components that receive a very low relative weight are discarded, or "die"; i.e., the number of samples drawn from them is so small that there is not enough information gained to perform another update. In the reference implementation of PMC by Kilbinger et al.

[Kil+11] that we use for the updating, the minimum number of samples per component is set at 20. We stop the update process when the convergence criteria of Algorithm 3 are met, and collect the sample used for inference in the final step. Note that we do not have to keep N constant in every step; in fact, we have experimented with reducing N as $N = K_{live} \times N_c$, where K_{live} is the number of live components. But we often saw PMC fail in those cases, as after a short number of steps, $K_{live} \to 1$ resulting in $\mathcal{P} \to 0$. Therefore, we recommend using identical values N in every PMC update step, but for the accuracy of inference, a larger sample is advisable in the final step.

4.3.1 Multimodal example revisited

Let us reconsider the 2D example with four modes, defined in Section 3.2, recalling MCMC's poor performance. The problem includes the LogGamma distribution in one of the two dimensions; for its asymmetry and its thick tail, LogGamma makes the example more realistic and more difficult for the proposal to adapt to. In particular, we avoid the accidental perfect fit of a Gaussian proposal to a Gaussian target. But note that the marginal $P(x_2)$ is a unit Gaussian, and the components in Fig. 4.5 capture this very well.

We run the new algorithm on the example with a relative suppression factor of $\omega = 1$ (all modes equally probable) and $\omega = 10^5$ (two modes suppressed). For completeness, the settings of the MCMC prerun, clustering, and PMC are listed in Table 4.1 in Row I and Row II respectively. The convergence statistics and the densities obtained with the final importance sampling step are displayed in Fig. 4.6.

For $\omega = 1$, all four modes have equal weight (cf. Fig. 4.6(b)), and the initial proposal at $t = 0$ produces a very good estimate of the target, with perplexity and ESS above 0.8. Convergence is achieved after two importance sampling steps, and the perplexity even rises in the final step $t = 2$; cf. Fig. 4.6(a). For $\omega = 10^5$, two of the four modes are strongly suppressed, and do not appear in the 2D marginal Fig. 4.6(d). In fact, the 6 components covering the suppressed modes "die" at $t = 1$ (Fig. 4.6(c)). After their removal, perplexity and ESS rise sharply by a factor of 2, reaching a similar level at $t = 1$ as for $\omega = 1$.

Running 20 chains for 40 000 iterations is a large computational investment for a 2D problem; the length is chosen large enough to be certain that the chains are able to explore the parameter space. However, when reducing the number of iterations to 5000, 1000, even to 500, we still obtain similarly good results as in Fig. 4.6: convergence after two steps, perplexity and ESS above 90 %; see Table 4.1, Row III. With only 500 iterations, the MCMC proposal is not adapted at all, showing that the initial proposal is good as the chains quickly move in on a target mode. If the initial MCMC proposal is not known to be this efficient, the chains should of course be run longer until successful adaptation.

In all cases, PMC converges very quickly and produces accurate results, demonstrating that our initialization procedure is successful. All modes are detected automatically in the prerun, and PMC properly determines the relative weights of the modes. Both perplexity and ESS are close to the maximum value of 100 %, showing excellent proposal adaptation to the target. Hence, the combination of MCMC and PMC succeeds where the local random walk MCMC by itself failed. After this successful proof of concept, we explore the impact of the target density on PMC in the following sections, and discuss how to adjust the various parameters to achieve satisfactory performance. It

Run	d	ω	N_{MCMC}	L	ν	N_c	K_g	N_g	t_{final}	$\mathcal{P}/\%$	ESS /%
I	2	1	40 000	100	12	500	3	4	2	96.5	94.7
II	2	10^5	40 000	100	12	500	3	4	2	96.0	93.7
III	2	1	≤ 5000	100	12	500	3	4	2	> 90	> 90
IV	22	10^5	50 000	150	12	500	22	4	5	56.6	22.9
V	22	10^5	50 000	150	-	1000	22	4	20	67.0	27.7
VI	22	10^5	50 000	150	12	500	32	4	6	61	44
VII	32	10^5	50 000	150	12	1000	12	1	-	-	-
VIII	42	10^5	100 000	300	12	2500	62	1	7	25	7.5
IX	12	-	50 000	200	12	500	62	1	4	69	50
X	32	-	50 000	200	-	1000	32	1	7	29	8.6

Table 4.1: Settings and results of various test runs referred to in the main text. d is the dimension of the target density's parameter space. ω is the suppression factor of two of the four maxima of the targets defined in (3.27) and (4.11). N_{MCMC} and L denote the length of a single chain and patch respectively. ν is the degree of freedom of each component in each $\mathcal{T}$ mixture proposal density, a missing value represents a $\mathcal{N}$ mixture. K_g is the number of components per group of chains, and N_c is the number of samples per component drawn during a PMC update step. N_g is the number of chain groups detected. t_{final} is the number of PMC updates before the final step, in which $\mathcal{P}$ and ESS characterize the quality of the adaptation of the proposal to the target. Common setting among all runs: in the MCMC prerun, the number of chains is $k = 20$, the Gaussian local random walk proposal function is updated after $N_{\text{update}} = 500$ iterations in $d = 2$ and $N_{\text{update}} = 1000$ iterations in $d > 2$. Chains are grouped according to a critical R value of $R_c = 1.2$. $N_{\text{final}} = 5 \times 10^5$ samples are collected in the final PMC step.

is instructive to go to extreme parameter values to see PMC's response, in particular when it fails. In turn, that knowledge serves as a "prior" to solve the inverse problem: where to tweak if PMC does not behave as desired? It turns out that the initialization is robust, and most issues are related directly to the more volatile importance sampling.

4.3.2 Higher dimensions

Having solved the 2D example for illustrative purposes, it remains to be be seen how far d can be increased until PMC suffers from the curse of dimensionality. In particular, we want to demonstrate that it is useful for the 30D global fit of Chapter 7, in the course of which we have developed the method. There exist many Monte Carlo approaches that only work in relatively few dimensions up to $d \lesssim 10$ (e.g. Appendix D) and we want to show that the combination of MCMC and PMC is more powerful.

To explore the effect of higher dimensions, we augment the 2D target density with an equal number of LogGamma and Gaussian distributions. Specifically, the target is

$$P(\boldsymbol{x}) = \prod_{i=1}^{d} P(x_i) \; , \tag{4.10}$$

where the distribution in the first two dimensions is given by (3.27) as before, and

$$P(x_i) = \begin{cases} \text{LogGamma}(x_i | l = 10, \lambda = 1, \alpha = 1), & 3 \leq i \leq \frac{d+2}{2} \\ \mathcal{N}(x_i | \mu = 10, \sigma = 1), & \frac{d+2}{2} < i \leq d \end{cases} \tag{4.11}$$

There are still four modes due to the first two dimensions; the extra dimensions do not add any further modes. For the purpose of drawing initial points for MCMC, we use a flat prior in all dimensions.

When d increases, the most important parameter to adjust is K_g, the number of components per group of chains. As a rule of thumb, K_g should be of the order of d, and slightly larger when $d > 15$. Another closely related issue is the number of samples per component, N_c, to be drawn in each importance sampling step. Evidently, as d increases, more samples are needed to explore the vicinity of each component, and ultimately, to better reduce the Kullback-Leibler divergence in the proposal update. This effect is a mild form of the curse of dimensionality: the number of parameters in the proposal, $\sim \dim \boldsymbol{\xi}$ (see (3.36)), is dominated by the scale matrices, and grows as d^2.

As a concrete example, consider $d = 22$, with $K_g = 22$ (all settings listed in Table 4.1, Row IV). The evolution of $\mathcal{P}$ and ESS is displayed in Fig. 4.7(a). In the first step, the perplexity is at a high level of 40 %, showing that the initialization with MCMC and hierarchical clustering works well. The PMC updates monotonously increase the perplexity as desired, but the ESS drops significantly at $t = 1$ and in the final step. These drops are caused by outliers with large importance weight, showing that the proposal function is not (yet) well adapted to the target. In particular, large weight outliers arise when the proposal does not have thicker tails than the target on the entire support; cf. Section 4.3.3. If the size of the importance sample is too small, such rare outliers may not even show up in every update step, but they are present in the larger final sample. Note that we collect only 44 000 samples for each update ($t < 5$), and 500 000 in the final step. So it happens that ESS seems stable in $t = 3$,4, convergence is declared, but in the final step, outliers are present and ESS drops by 60 %.

We repeat the run with $N_c = 1000$ using Gaussian components, that is the thinnest tails available to us; cf. Fig. 4.7(b) and Table 4.1, Row V. Now the perplexity is pretty stable at an even higher level than with $\mathcal{T}$ components, but ESS shows enormous volatility, and convergence — the stability of perplexity *and* ESS over two steps within 2 % — cannot be declared until the run is aborted at $t_{max} = 20$. This is not to say that the final result is unusable, but it leads us to discourage the use of ESS for convergence monitoring with PMC. After all, PMC explicitly attempts to minimize the Kullback-Leibler divergence, so $\mathrm{KL}(P\|q) \searrow 0 \Rightarrow \mathcal{P} \nearrow 1$. Hence if PMC cannot increase $\mathcal{P}$, further updates are unnecessary. Considering only $\mathcal{P}$, convergence is declared after four steps in this example, and the final result is of similar quality as in the run with $\mathcal{T}$ proposal. Using a more computing-intensive run with a $\mathcal{T}$ mixture with ten additional components and more samples per component compared to the Gaussian run V, we observe that the perplexity and ESS rise monotonously (details in Table 4.1, Row VI). Hence outliers do not appreciably affect PMC in that case, and the ESS increases by (60 – 90) %, mostly due to extra components covering the tails more accurately. $\mathcal{P}$ rises by only 5 % with the extra components, but even then $\mathcal{P}$ is still 5 % less than in the Gaussian case. The Gaussian proposal describes the Gaussian directions in P very well, and thus results in a better average weight.

It is surprising to see what happens when K_g and N_c are too small. The PMC update algorithm then displays a “suicidal” tendency to assign low weights to many components. As a consequence, these components “die out”, as no more samples are drawn from them. With more updates, the rate of component deaths grows, until finally there is only one component left; PMC has failed to adjust the proposal, and the importance weights cannot be used for inference. An example is shown in Fig. 4.8(a), where we focus on a single mode in $d = 32$ dimensions for simplicity (Table 4.1, Row VII). The number of components is stable at 12 until $t = 6$, then drops quickly until only one is left at $t = 15$. At the same time, perplexity and ESS tend to zero. A characteristic of PMC failing is that the sizes of components shrink far too much; e.g, the $(1, 1)$ element of the surviving component's scale matrix $\boldsymbol{\Sigma}$ decreases by a factor of 30 from $t = 5$ to $t = 18$.

Note that the fact that components *can* die has some desirable ramifications. In our example with $\omega = 10^5$, all components in the suppressed modes die out in the first proposal update, because their contribution to the integral is negligible. Unfortunately, if components are not placed close enough to a region of high probability, they face the same situation as if they were in a suppressed mode. The Kullback-Leibler divergence is usually a complicated function with many local minima, few of which represent an efficient adaptation of q to P. A variant of expectation-maximization (Appendix E), PMC updates tend toward a *local* mode. Cf. [Cap+08] for an instructive example where KL is known to have three minima, one of which leads to a bad fit with components dying out. This highlights the crucial impact of choosing a good initial proposal and provides the motivation to develop the method presented here.

We verified that PMC works up to $d = 42$; see Table 4.1, Row VIII and cf. Fig. 4.8(b). At this large dimension, the fit fails unless a massive number of $K_g = 62$ components and $N_c = 2500$ samples per component are used. Nevertheless, $\mathcal{P}$ and ESS are much lower than at $d = 22$ — evidently a sign of the curse of dimensionality. The update procedure itself — computing new component weights, means, and scale matrices — consumes a significant amount of time, about one minute on a single core of an Intel i7-2600 operating at 3.4 GHz, whereas the importance sampling (including input and output) takes only 2 s. While it is not difficult to thread-parallelize the update to increase the speed, it becomes nevertheless apparent that PMC hits its limits of applicability at $d \approx 40$. Unless the target happens to be a Gaussian or Student's t, the task of adjusting the proposal to match the target, in particular in the tails, becomes increasingly difficult, as witnessed by low values of $\mathcal{P}$ and ESS in the final step. Outliers can affect 1D and 2D marginal distributions; cf. Fig. 4.10(b).

4.3.3 Degeneracy

In real-life problems, the formulation of the statistical model often contains parameters that provide a redundant explanation of the data. The posterior value is (nearly) constant along a connected subregion. Such a "flat direction" is called a *degeneracy*. We simulate this situation by adding two flat directions to the $(+,+)$ mode of the previous example target $P(x)$ defined in (4.10) and (4.11). Specifically, our new, unimodal target, $P_f(\boldsymbol{x})$, factorizes as

$$P_f(x_i) \propto \begin{cases} \text{LogGamma}(x_i|10,1,1), & i = 1 \vee 3 \leq i \leq \frac{d}{2} \\ \mathcal{N}(x_i|10,1), & i = 2 \vee \frac{d}{2} < i \leq d-2 \\ \text{const}, & d-1 \leq i \leq d\,. \end{cases} \tag{4.12}$$

Wraith et al. [Wra+09] quote the ability to deal with degeneracies as one of the strengths of PMC; in the following, we want to verify this. Let us first consider P_f for $d = 12$. Using the settings in Table 4.1, Row IX, the initial components are properly distributed along the flat direction (Fig. 4.9(a)) and give a high starting perplexity of 63 %. With just three PMC updates, the final step importance sample yields good results: $\mathcal{P} = 69\,\%$ and ESS = 50 %. The marginal distribution $P(x_1, x_{12})$ is shown as a histogram in Fig. 4.9(b). We note that the region of maximum probability around $x_1 = 10$ is not covered very accurately, in particular the bins with the maximum weight, colored in red, are few and scattered far apart. These bins come from samples with an importance weight w far above average. As a reminder, large $w = P(\boldsymbol{x})/q(\boldsymbol{x})$ usually appears when the tail of the target is thicker than the proposal's; a situation that is much more likely to happen with degeneracy.

The question arises how to reduce the effect of isolated histogram bins with large weight. We seek an alternative nonparametric *density* estimation based on the discrete set of samples — 500 000 $(\boldsymbol{x}, w)$ pairs in this case. Our choice is kernel density estimation (KDE) for its smoothing capabilities; details on KDE are presented in Appendix F. The KDE-smoothed output makes it easier to quickly grasp the important structures of the underlying density due to its visual appeal; cf. the clear vertical band in Fig. 4.9(c). Its virtues in computing marginal distributions are discussed below, and exploited in the global fit results shown in Section 7.2.

4.3.3.1 Cropping

In higher dimensions, importance sampling suffers much more from outliers with very large relative importance weights. To illustrate this, let us use an ill-fitting Gaussian (instead of a $\nu = 12$ Student's t) mixture in $d = 32$ (Table 4.1, Row X). Ignoring the oscillating ESS, the run converges after 7 steps, with $\mathcal{P} = 0.29$ and ESS = 0.09 in the final sample. The distribution of the 500 000 final-step importance weights is shown in Fig. 4.10(a) on a logarithmic scale of the ordinate spanning more than five orders of magnitude. Remarkably, nearly all weights are in the first bin, only ~ 100 samples, the outliers, have a larger weight. In the 2D plot Fig. 4.10(b), the single event with the largest self-normalized weight of $\bar{w} = 5.2\,\%$ dominates completely, and the vertical band at $x_1 = 10$ is barely visible despite the KDE smoothing applied. This behavior is highly undesirable. There are several remedies: we could salvage this simple example by choosing a proposal function with fat tails and more components, but mild outliers would still be present. In general, if the target's shape is sufficiently complicated, the PMC update is not able to guarantee good covering properties in the tails. It appears outliers are inevitable for $d \gtrsim 25$ — a major weak point of importance sampling.

A practical approach to get the most out of a given set of importance samples marred by outliers is to simply remove the outliers. Cropping the 200 samples with highest weight in the previous example leads to Fig. 4.10(c). The result is exactly what we expect, and in very good agreement with the output in the simpler 12D case of Fig. 4.9(c), which is and should be identical up to Monte Carlo uncertainty. As a cross check, we compare the contours of the marginal $P(x_1, x_{32})$ produced by MCMC and the filtered PMC output (not shown) and again find good agreement. In addition, it is useful to consider the value of the target density in the removed samples. While the maximum target value in the entire importance sample is $\max \log P = -163.6$, we find a maximum (minimum) log value of −173.3 (−184.7) among the filtered samples. This confirms that the outliers occur out in the tails of $P(\boldsymbol{x})$, and that we do not miss local modes by removing them.

4.4 Short guide to parameter settings

At this point, we summarize the previous sections to provide guidance on the various tunable parameters to a novice that wants to use PMC with MCMC initialization. The knowledge comes from running the examples described in this chapter, but also from the experience with the global fit in Chapter 7. Crucial settings of the particular runs are listed in Table 4.1 and Table 7.4.

For the MCMC step, we used $k = 20 \dots 50$ chains to discover all four modes of the example target density. The chains were run for 10 000 ($d \gtrsim 2$) – 100 000 ($d = 42$) iterations with a Gaussian proposal, though Student's t could be used as well. Discarding the initial 20 % for burn-in, we split up the chains into patches of length $L = 100 - 300$, the exact value of L is not critical. Patches during which no move is accepted are discarded, those where the numerical Cholesky decomposition fails are used with off-diagonal elements set to zero.

With regard to hierarchical clustering, we group chains according to the R values, using a threshold value of $R_c = 1.1 - 2$. The number of components per group, K_g, ought to be $\gtrsim d$; the bigger K_g, the more accuracy is obtained at the expense of more evaluations of the target. The initial components arise from long patches of chains within a group. Hierarchical clustering is stopped if the distance measure in two consecutive steps is reduced by less than $\varepsilon_{min} = 10^{-4}$.

In the PMC step, we initially set all component weights equal. In most applications, a Gaussian mixture has tails that are thinner than the target's tails, so one can decide for a Student's t mixture with degree of freedom $\nu = 2 \dots 15$. Good results were obtained in the examples with $N_c = 500$ ($d = 2$), 1000 ($d = 12$), and 2500 ($d = 42$) samples per component. Convergence is declared when the normalized perplexity $\mathcal{P}$ is stable to within 2 % between consecutive steps, or when it exceeds 92 %. Jumps in the ESS hint at outliers caused by too few mixture components or by a proposal whose tails are too thin. If the PMC updates "kill" more and more components and reduce the perplexity, more initial components and a larger sample size may help. If outliers have a dominant effect on the resulting marginal distributions, the combined effect of KDE smoothing and outlier removal provides a partial remedy. After convergence, a sample size of 500 000 is adequate when the $\mathcal{P} \lesssim 1$ and d is small, and a size ranging in the millions is recommended for targets in higher dimensions where $\mathcal{P}$ is small.

4.5 Outlook

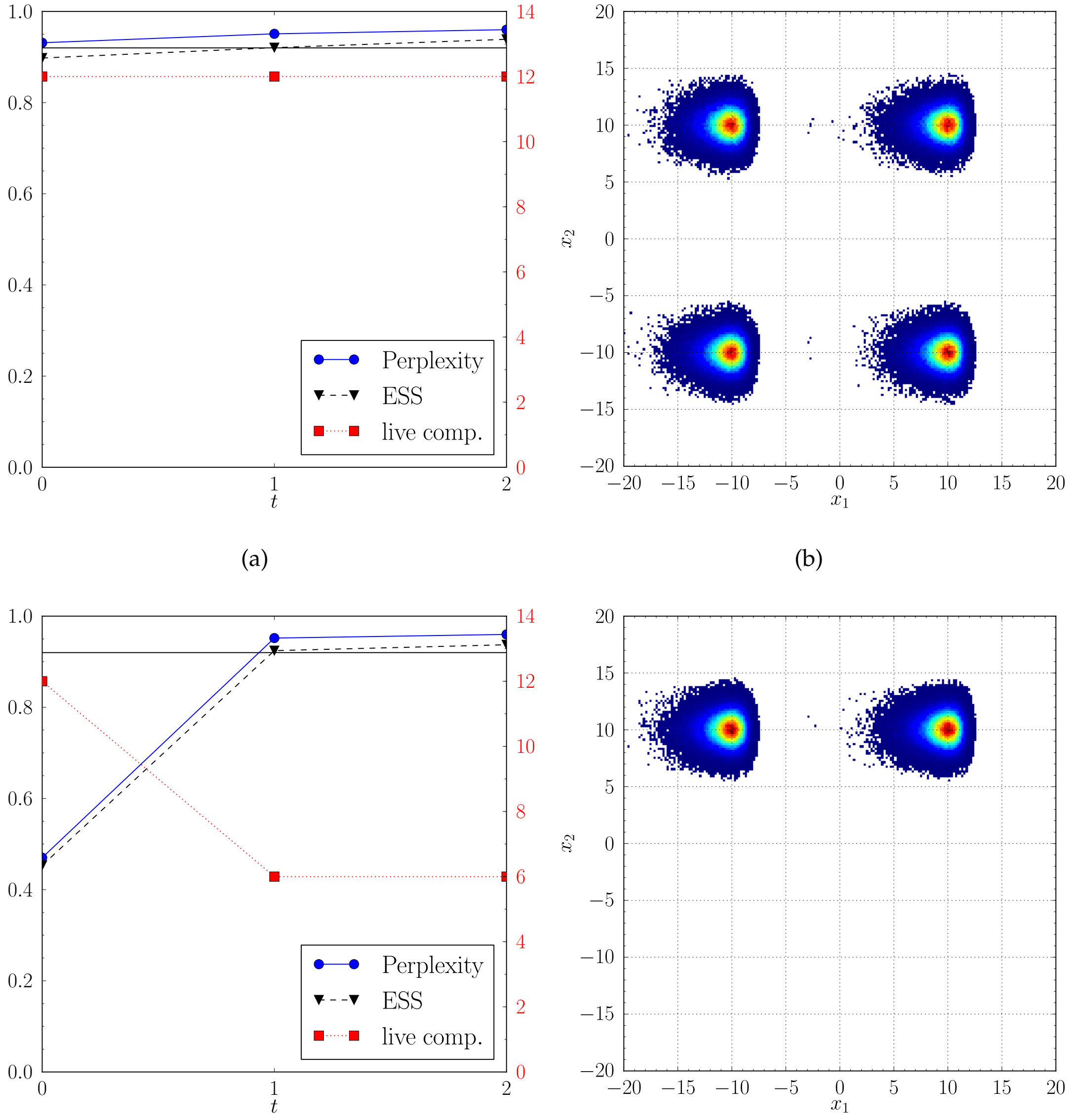
We presented a new method that cleverly initializes adaptive importance sampling to replace manually inputting knowledge of the target density. The components of the proposal mixture density are extracted from a MCMC prerun with the help of hierarchical clustering. The support of multimodal and degenerate targets is reliably found and well covered by the proposal, allowing quick and successful adaption of the proposal via PMC. In the development of the method, our focus was to successfully perform the complicated 30D global fit described in Chapter 7. Having achieved that, we now discuss the potential for future improvement.

Our examples suggest that importance sampling works up to $d \approx 40$, but problems with outliers appear already for $d \gtrsim 20$. To a certain degree, they can be circumvented by more mixture components, an adjustment of the Student's t degree of freedom ν, and a larger sample size. We presented guidance how to manually adjust these parameters, but an automatic adjustment is highly preferred. [HOVD11] show that ν can be updated along with the component weight, means, and variances. The soft limit of $d \approx 40$ arises because importance sampling starts to falter, but the MCMC initialization still works well. It is thus conceivable that marginal distributions are less affected by outliers if the proposal function of the final PMC step is used as a global proposal in MCMC. In the spirit of the global-local proposal (Appendix D.1), individual mixture components guide local jumps, and the full proposal is used for global jumps. Using a fixed proposal and assuming rapid mixing due to the global jumps, massive parallelization is straightforward — individual chains need not run for 1000's of iterations before the results are usable.

Minor improvements are possible when replacing the Gaussian clustering with the considerably more involved Student's t clustering to obtain a Student's t mixture proposal density from the chain patches [EAPG09]. However, the more urgent problem is to determine the number of mixture components from the prerun; the Bayesian or Akaike information criteria [Sch78; Aka74] may prove useful, and could eliminate the need for chain grouping. A promising alternative to hierarchical clustering is the *variational Bayes* approach described in [BGP10], in which the "best" number of components is computed along with the positions and covariances of the reduced mixture's components.

At a more fundamental level, one could eliminate the MCMC prerun entirely in favor of a large number of samples from the prior or a uniform distribution on the parameter space. Two advantages are that the samples can be computed with massive parallelization and that potentially fewer samples are required by avoiding the redundancy of multiple chains in the same region. One would need to replace hierarchical clustering with a more sophisticated algorithm that explicitly takes into account the value of the target at each sample point — this information is not used at present. Ideally, it should also determine the number of mixture components — similar to the initial stage of nested sampling [Ski06; FHB09]. But by giving up the MCMC prerun, we suspect there is a greater chance that suppressed modes and degeneracies are missed or poorly captured in high-dimensional problems. In addition, it proved useful for validation purposes to compare the marginal distributions from MCMC and PMC for qualitative agreement. If a region is visible in the MCMC but not in the PMC output, either PMC failed, or that region contains negligible probability mass, which can be verified with the samples' target values in that region.

During the final stages of preparing this work, we became aware of a recent effort that comprises many of the above points. Cornuet et al. [Cor+12] use a large sample from the uniform or prior distribution with a logistic rescaling to learn the features of the target. They run Gaussian mixture clustering with the integrated likelihood criterion determining the optimal number of components of the initial proposal for PMC. By cleverly combining the samples of *all* PMC update steps, and not only the most recent one as in our approach, they report a significant Monte Carlo variance reduction.



(a) (b)

(c) (d)

Figure 4.6: The convergence diagnostics (a), (c) and posterior density (b), (d) resulting from PMC runs I and II; cf. Table 4.1. The maxima with negative x_2 are suppressed by a factor of 1 (upper panels) and 10^5 (lower panels). Note that three components per mode are used in step 0 in both runs. The stopping criterion of 0.92 for perplexity and ESS is indicated by the horizontal line in (a) and (c). Compare (b) and (d) with clustering output in Fig. 4.5(b).

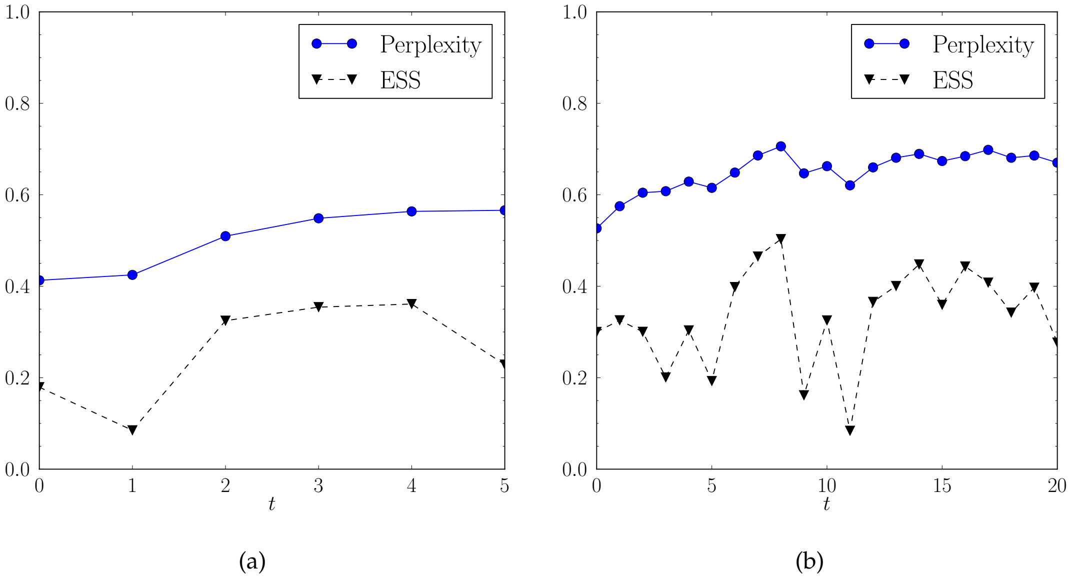


Figure 4.7: Evolution of PMC convergence criteria for the runs IV and V of the example target in $d = 22$ with (a) Student's t and (b) Gaussian components. Outliers cause ESS to drop, especially in the final step with sample size 500 000.

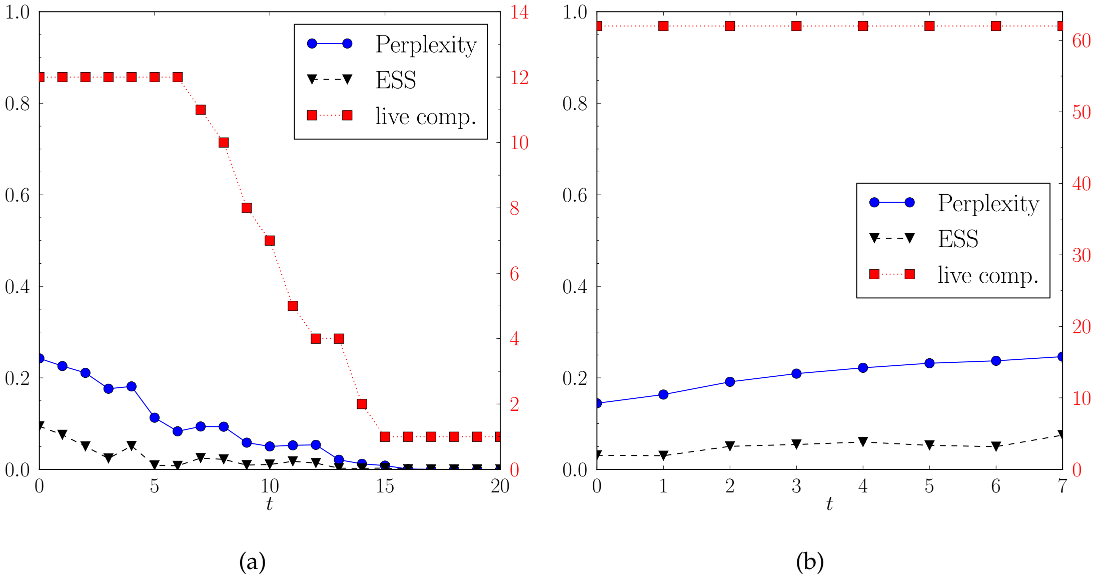
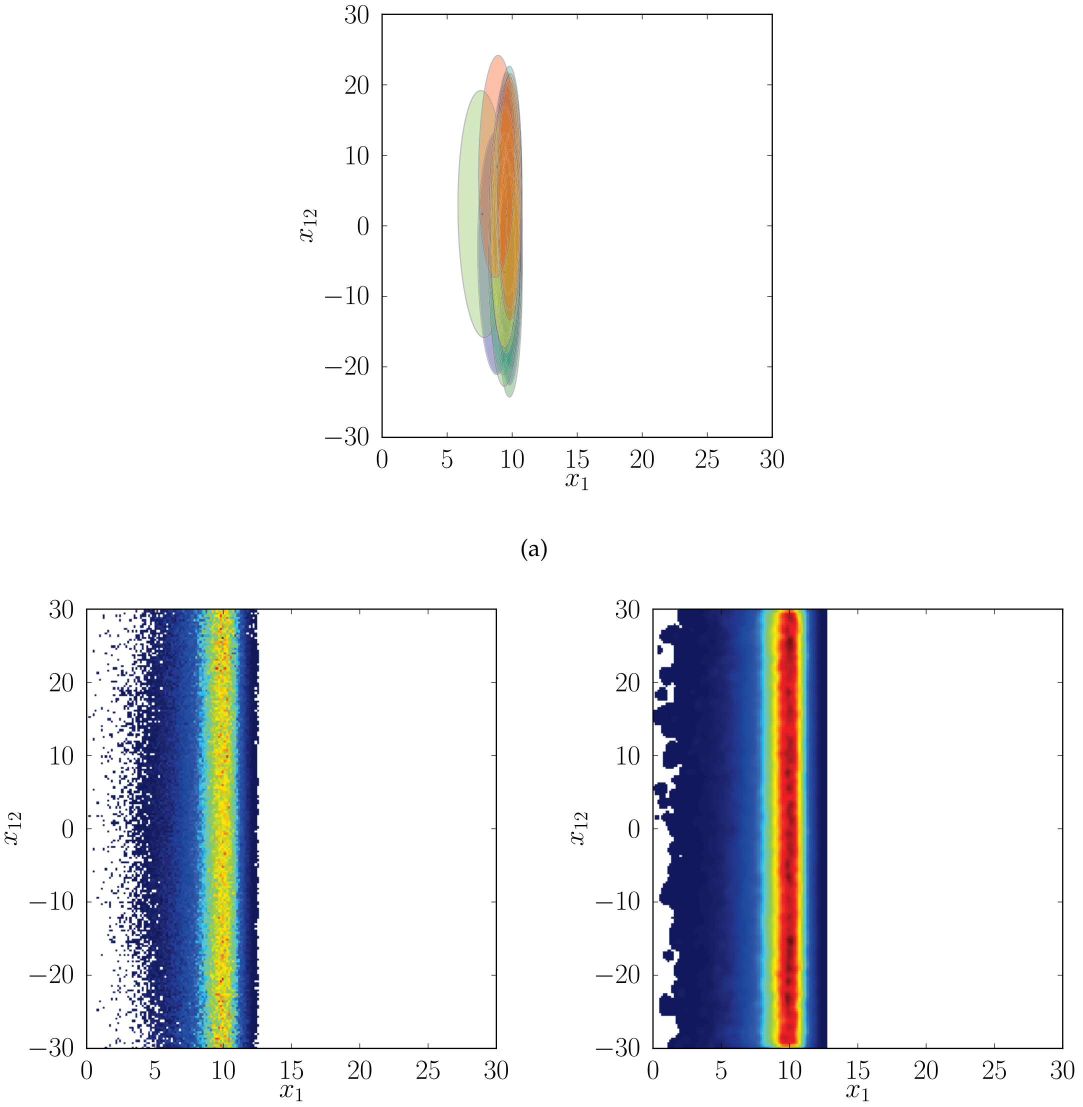


Figure 4.8: Evolution of PMC convergence criteria for the runs VII and VIII with unimodal targets. (a) $d = 32$ dimensions with only $K_g = 12$ components. PMC fails, components start to die after the update at $t = 6$; perplexity and ESS drop to zero. (b) $d = 42$ dimensions with $K_g = 62$ components. PMC monotonously increases $\mathcal{P}$, and converges at a low level. All components stay alive.

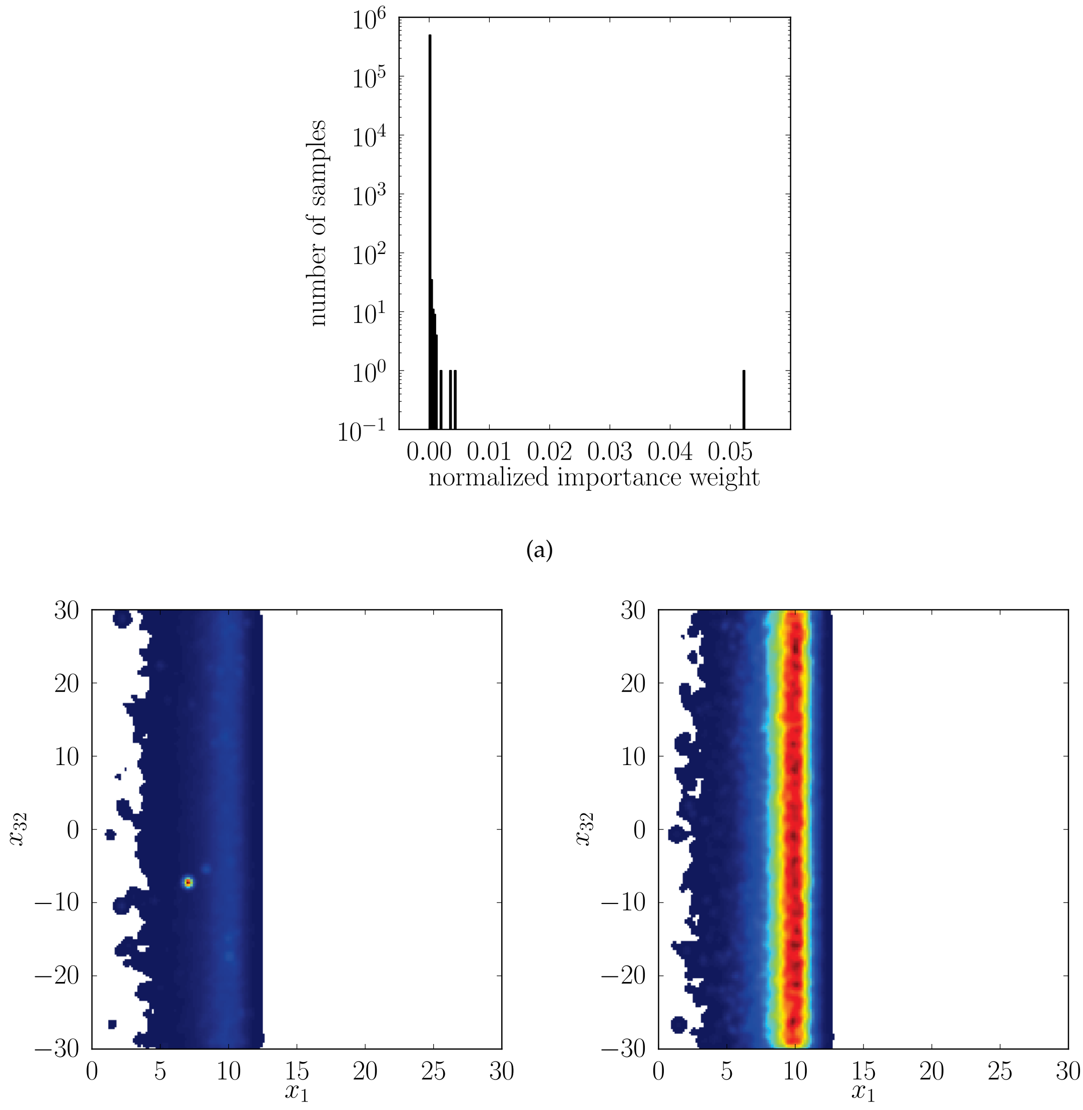


(a)

(b)

(c)

Figure 4.9: Results of PMC run IX for the marginal distribution $P_f(x_1, x_{12})$ in $d = 12$. The 1D marginal distributions are $P(x_1) = \text{LogGamma}(x_1|10, 1, 1)$ and $P(x_{12}) = \text{const.}$ (a) $1\,\sigma$ ellipses of the components of the *initial* mixture proposal density. (b) Histogram approximation. (c) KDE with pixels whitened with an intensity of 10^{-4} less than the maximum.



(a)

(b)

(c)

Figure 4.10: Results of PMC run X for the marginal distribution $P_f(x_1, x_{32})$ in $d = 32$ using a poorly fitting Gaussian proposal. The 1D marginal distributions are $P(x_1) = \text{LogGamma}(x_1|10,1,1)$ and $P(x_{32}) = \text{const.}$ (a) Histogram of importance weights with logarithmic scale discloses outliers. The dominant outlier at $(x_1, x_{32}) = (7.0, -7.3)$ has a weight of $\bar{w} = 5.2\,\%$ and dominates the KDE (b) of $P_f(x_1, x_{32})$. But when the 200 highest-weight samples are cropped, the KDE (c) is in very good agreement with $P_f(x_1, x_{12})$ of Fig. 4.9(c).

5 Theory of rare *B* decays

In this chapter, we want to establish the notation and the theoretical basis for describing rare B decays that we use in the search for new physics. After reviewing the standard model of particle physics with emphasis on the quark mixing in Section 5.1, we introduce a model-independent approach, the $\Delta B = 1$ effective field theory (EFT), to describe the decays quantitatively in Section 5.2. Finally, the main difficulties on the theory side — the nonperturbative effects of QCD — are briefly discussed in Section 5.3 Throughout, we follow [Bea+12; Dyk12] and work in natural units where $c = \hbar = 1$.

5.1 Standard model

The standard model (SM) of particle physics, also known as the Glashow-Salam-Weinberg model, is a quantum field theory that describes the strong, weak, and electromagnetic forces in terms of local gauge symmetries; the SM gauge group is

$$SU(3)_{\mathrm{C}} \times SU(2)_{\mathrm{L}} \times U(1)_{\mathrm{Y}} \,. \tag{5.1}$$

In compact notation, the Lagrangian density of the SM is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{SM}} = & \imath\bar{\psi} \not{D} \psi - \frac{1}{4}\left[A^a_{\mu\nu}A^{\mu\nu}_a + B_{\mu\nu}B^{\mu\nu} + G^b_{\mu\nu}G^{\mu\nu}_b\right] \\ & - \bar{\psi}_L \boldsymbol{Y} \phi \psi_R + (D_\mu\phi)^*(D^\mu\phi) - V(\phi) + \mathrm{h.c.}\end{aligned} \tag{5.2}$$

$\mathcal{L}_{\mathrm{SM}}$ describes fermions — quarks and leptons — with Dirac spinors ψ and their interactions mediated by the gauge fields through the respective field strength tensors $A^a_{\mu\nu}$ ($SU(2)_{\mathrm{L}}$), $B_{\mu\nu}$ ($U(1)_{\mathrm{Y}}$), and $G^b_{\mu\nu}$ ($SU(3)_{\mathrm{C}}$). The gauge fields appear in the covariant derivative D and the gauge kinetic terms. The symmetry group $SU(2)_{\mathrm{L}} \times U(1)_{\mathrm{Y}}$ is spontaneously broken through the Higgs mechanism [EB64; Hig64; GHK64]. ϕ is the $SU(2)_L$ Higgs doublet, coupled to the fermions through (matrix-valued) Yukawa couplings $\boldsymbol{Y}$; the Higgs potential is

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4, \ \mu^2 > 0 \,. \tag{5.3}$$

$\psi_{L/R}$ denotes a left- [right-] chiral Dirac spinor. We omit any gauge-fixing, ghost, and counter terms needed to render S matrix elements of physical processes finite in the quantized field theory.

The fermion content of the SM consists of three generations of leptons and quarks; left-handed particles are combined into $SU(2)_L$ doublets. In the lepton sector, the doublets are composed of charged leptons and neutrinos; e.g., the first generation contains the electron e and the electron neutrino ν_e. In the quark sector, doublets contain pairs of one up- and one down-type quark. The right-handed charged leptons and quarks are singlets under $SU(2)_L$, and the right-handed neutrinos are not part of the SM. Local

$m_{u,d}$	$m_{c,s}$	$m_{t,b}$
~ 0.003	~ 1.3	~ 170
~ 0.005	~ 0.1	~ 4.2

Table 5.1: Quark masses in GeV [Nak+10]. Each column contains one quark generation. The numbers are renormalization scheme dependent, but purport the proper order of magnitude.

$SU(2)_L$ gauge invariance does not permit mass terms for leptons and quarks, since the product $\bar{\psi}_L \psi_R$ is not gauge invariant.

The gauge groups contribute a number of vector bosons: there are eight massless gluons associated with the unbroken color group $SU(3)_\mathrm{C}$ of the strong interaction, and the broken electroweak $SU(2)_\mathrm{L} \times U(1)_\mathrm{Y}$ symmetry yields the three massive vector bosons of the weak force, the $W^\pm$ and the Z, and the massless photon γ of electromagnetism. Last, the standard model predicts the existence of a massive, neutral, scalar particle, the *Higgs* boson H, corresponding to one of the four degrees of freedom of the complex $SU(2)_L$ doublet ϕ that is needed for the spontaneous breaking of $SU(2)_\mathrm{L} \times U(1)_\mathrm{Y}$ and for mass terms of quarks and leptons from the Yukawa interactions, see Section 5.1.1. Note that the neutrinos remain massless.

After decades of searches for the Higgs boson, the LHC experiments ATLAS and CMS have reported strong evidence of a new boson with a mass around 125 GeV at the time of writing [Aad+12a; Cha+12a]. If that particle indeed is the Higgs boson, then at last all SM particles have been observed, completing the enormous success of the SM.

5.1.1 Quark mixing

The three generations of quarks have identical gauge properties; the only distinction is the quark mass. The u, d, and s quarks are nearly massless, but the c, b and t quark masses exhibit a clear mass hierarchy; cf. Table 5.1. It is useful to introduce the *flavor* quantum number for each quark, such that, e.g., a b quark has bottomness[1] $B = -1$, and its antiparticle $\bar{b}$ has bottomness $B = +1$. Mass eigenstates (u, d) do not coincide with the interaction eigenstates $(\tilde{u}, \tilde{d})$; rather, they are connected by the 3×3 Yukawa matrices $\boldsymbol{Y}_u$ and $\boldsymbol{Y}_d$ for up- and down-type quarks. When the Higgs field ϕ condenses, it acquires a vacuum expectation value $\langle \phi \rangle = (0, \langle H \rangle)$ with $v = \sqrt{2}\,\langle H \rangle \approx 246\,\mathrm{GeV}$ [Ber+12] and produces mass terms for the quarks, leptons, and the W and Z bosons. In the interaction basis, in which covariant derivatives are flavor-diagonal, the Yukawa terms for the quarks $\tilde{q}$ then become

$$\begin{aligned}
\mathcal{L}_\mathrm{SM} \supset v \left[\bar{\tilde{u}}_L \boldsymbol{Y}_u \tilde{u}_R + \bar{\tilde{d}}_L \boldsymbol{Y}_d \tilde{d}_R \right] && (5.4)\\
= v \left[\bar{\tilde{u}}_L \boldsymbol{V}_{1,u} \boldsymbol{M}_u \boldsymbol{V}^\dagger_{2,u} \tilde{u}_R + \bar{\tilde{d}}_L \boldsymbol{V}_{1,d} \boldsymbol{M}_d \boldsymbol{V}^\dagger_{2,\mathrm{d}} \tilde{d}_R \right] && (5.5)\\
\equiv v \left[\bar{u}_L \boldsymbol{M}_u u_R + \bar{d}_L \boldsymbol{M}_d d_R \right] \;, && (5.6)
\end{aligned}$$

where we changed to the mass basis by diagonalizing $\boldsymbol{Y}_u$ and $\boldsymbol{Y}_d$ with the unitary matrices $\boldsymbol{V}_{1,u}, \boldsymbol{V}_{1,d}$, $\boldsymbol{V}_{2,u}$, and $\boldsymbol{V}_{2,d}$ [Nak+10, Ch. 11]. The masses are related to the

[1] We follow the slightly misleading convention of using the same symbol B to denote both the B meson and the quantum number bottomness.

diagonal matrices, $\boldsymbol{M_u}$ and $\boldsymbol{M_d}$, and the Higgs vacuum expectation value by

$$v\boldsymbol{M}_u = \mathrm{diag}(m_u, m_c, m_t), \qquad v\boldsymbol{M}_d = \mathrm{diag}(m_d, m_s, m_b)\ . \tag{5.7}$$

In the remainder of this work, we are concerned with quark flavor *changes*. In the SM, the electromagnetic and the strong force conserve flavor. The only interaction term that mixes quark flavors is the charged current involving the weak bosons $W^\pm$

$$\mathcal{L} \supset \mathcal{L}_{\mathrm{CC}} = -\frac{g}{\sqrt{2}}(\boldsymbol{V}_{1,u}\boldsymbol{V}^\dagger_{1,d})_{ij}\bar{u}_i W^+_\mu \gamma^\mu P_L d_j + \mathrm{h.c.} \tag{5.8}$$

where g is the coupling constant of $SU(2)_{\mathrm{L}}$, $i = u, c, t$ denotes the up-quark generation, $j = d, s, b$ represents the down-type generations, γ^μ is a Dirac matrix, and P_L is the left-chiral projector, implementing the $V - A$ structure of maximum parity violation; i.e., only left-chiral quarks couple to the $W^\pm$ bosons. The magnitude of flavor mixing between an up-type quark i and a down-type quark j is given by the matrix element

$$V_{ij} \equiv (\boldsymbol{V}_{1,u}\boldsymbol{V}^\dagger_{1,d})_{ij} \tag{5.9}$$

of the celebrated Cabibbo-Kobayashi-Maskawa (CKM) matrix [Cab63; KM73]

$$\boldsymbol{V}_{\mathrm{CKM}} \equiv \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}\ . \tag{5.10}$$

$\boldsymbol{V}_{\mathrm{CKM}}$ is a unitary matrix with four real degrees of freedom. It is well confirmed experimentally that mixing across generations is strongly suppressed. In this work, we will therefore use the Wolfenstein parametrization [Wol83], which makes this hierarchy explicit. Using the four parameters $A \approx 0.8, \lambda \approx 0.2, \bar{\rho} \approx 0.1$, and $\bar{\eta} \approx 0.4$ (more accurate values are given in Table A.5), we find

$$\boldsymbol{V}_{\mathrm{CKM}} = \begin{pmatrix} 1 - \lambda^2/2 & \lambda & A\lambda^3(\bar{\rho} - i\bar{\eta}) \\ -\lambda & 1 - \lambda^2/2 & A\lambda^2 \\ A\lambda^3(1 - \bar{\rho} - i\bar{\eta}) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}\left(\lambda^4\right)\ , \tag{5.11}$$

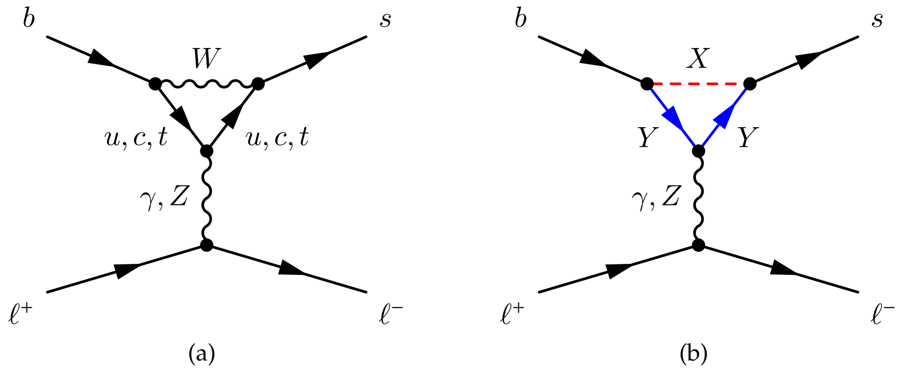
and $\boldsymbol{V}_{\mathrm{CKM}}$ is unitary to all orders in λ in this parametrization [Nak+10, Ch. 11].

Due to the unitarity of the CKM matrix, there are six conditions on the elements V_{ij}, which can be visualized as *unitarity triangles* in the suitable complex plane. Any deviation from unitarity indicates NP. Two collaborations, CKMfitter [Cha+05, frequentist] and UTfit [Bon+06, Bayesian] endeavor to extract the four CKM parameters from a large number of decays. Results are available for explicit models, such as the SM or various NP models, and for the generic class of models that include only SM tree-level processes. For our purposes, we will only need the following triangle:

$$0 = V_{ub}V^*_{us} + V_{cb}V^*_{cs} + V_{tb}V^*_{ts}\ . \tag{5.12}$$

5.1.2 Flavor changing neutral currents

There are no FCNCs at tree level in the SM; i.e., there is no interaction vertex linking different quark flavors of the same electric charge ($-1/3$ or $2/3$). The only possible



(a) (b)

Figure 5.1: Sample FCNC Feynman diagram contributions that describe $b \to s\ell^+\ell^-$ decays (a) in the SM and (b) with NP particles X and Y.

transition is via loop processes involving two charged-current interactions from the term (5.8); cf. Fig. 5.1(a) for an example Feynman diagram.

The reason why FCNCs are interesting is that transitions occurring only through FCNCs are strongly suppressed in the SM. Therefore, any NP contribution with unknown heavy particles in the loop as shown in Fig. 5.1(b) could a-priori be of the same order of magnitude as the SM contribution. Essentially, all precise measurements of tree-level processes confirmed the SM in the sense that all flavor changes are mediated by $\boldsymbol{V}_{\mathrm{CKM}}$ and ruled out sizable NP contributions; cf. the continuously updated analyses of the CKMfitter [Cha+05] and UTfit [Bon+06] collaborations. But for FCNC-mediated processes, the experiments have not yet reached such a level of accuracy. FCNC decays are thus a prime target of investigation, potentially showing hints of new physics at an energy scale much higher than that of the decaying particle b. Thus one can probe energy scales in the TeV range with a particle collider operating at 10 GeV; cf. Section 6.2.1 on the B-factory experiment Belle.

We consider the $b \to s$ transition as our main example to illustrate the mechanisms responsible for the suppression of FCNC decays in the SM. The generic amplitude is

$$\mathcal{A}_{b\to s} = V_{ub}V_{us}^* f(\hat{m}_u^2) + V_{cb}V_{cs}^* f(\hat{m}_c^2) + V_{tb}V_{ts}^* f(\hat{m}_t^2)\,, \qquad \hat{m}_q \equiv m_q/m_W\,, \tag{5.13}$$

with a process-dependent loop function $f(\hat{m}^2)$ containing a prefactor of $(g/4\pi)^2 \ll 1$, an example of *loop suppression* for a small coupling constant. Using the unitarity triangle (5.12), we remove V_{cj} from the amplitude (5.13):

$$\mathcal{A}_{b\to s} = V_{tb}V_{ts}^* \left[f(\hat{m}_t^2) - f(\hat{m}_c^2) \right] + V_{ub}V_{us}^* \left[f(\hat{m}_u^2) - f(\hat{m}_c^2) \right] \,. \tag{5.14}$$

The first term in $\mathcal{A}_{b\to s}$ is mildly *CKM suppressed*, for $V_{tb}V_{ts}^* \propto \lambda^2 \simeq \mathcal{O}\left(10^{-2}\right)$. The second term is further CKM suppressed with respect to the first, as $V_{ub}V_{us}^*/V_{tb}V_{ts}^* \simeq \mathcal{O}\left(10^{-2}\right)$.

The quark mass hierarchy (cf. Table 5.1) plays an important role here; an expansion of the loop functions yields $f(\hat{m}_u^2) - f(\hat{m}_c^2) \simeq \mathcal{O}\left(10^{-4}\right)$, a result of the Glashow-Iliopoulos-Maiani (GIM) mechanism [GIM70]. The same suppression is in effect for any pair of quarks $i \neq j$ such that both $\hat{m}_i^2$ and $\hat{m}_j^2$ are $\ll 1$. Hence, the only unsuppressed contributions are those involving a top quark like the first term in (5.14), because only $\hat{m}_t^2 > 1$. In numbers, the SM branching ratios for $b \to s$ transitions range from $\mathcal{O}\left(10^{-4}\right)$ $(B \to K^*\gamma)$

over $\mathcal{O}\left(10^{-7}\right)$ $(B \to K^{*}\ell^{+}\ell^{-})$ to $\mathcal{O}\left(10^{-9}\right)$ $(B_s \to \mu^{+}\mu^{-})$; hence they are rare decays, but still within experimental reach. In contrast, the decays with the equivalent up-type, quark-level transition $t \to c$ are at the level of $\mathcal{O}\left(10^{-12}\right) - \mathcal{O}\left(10^{-14}\right)$ in the SM [Agu04], and thus not detectable with present or even next-generation experiments. Within the *minimal supersymmetric extension of the SM* (MSSM), rates can be enhanced, but remain too small to be detected [Beh+12]. Similarly, we do not consider the $\Delta B = 2$ transition $b \to d$ in our fit. First because of extra CKM suppression $|V_{td}/V_{ts}| \approx 10^{-2}$, and second, because it could only constrain the Wilson coefficients of the $\Delta B = 1$ EFT under the assumption of *minimal flavor violation*. Therefore, we focus on $b \to s$ only.

Comparing $\boldsymbol{V}_{\mathrm{CKM}}$ (5.9) with the Wolfenstein parametrization (5.11), we observe that in (5.14), only the second term has a nonzero imaginary part, and is thus responsible for CP violation in the SM; but if we consider only CP conserving observables, we may safely ignore that term due to its small relative contribution.

5.2 Effective field theory

In quantum field theory, the prediction of an observable reaction with *real* particles requires including the effects of all *virtual* particles, where the latter typically appear at much higher energy scales than the former. The concept of an *effective field theory* (EFT) provides a framework to simplify the multiscale problem by reducing it into separate single-scale problems. A Lagrangian is constructed with fields describing the relevant degrees of freedom at the low scale, and the effects of particles that only appear at the high scale are absorbed in the coupling constants of the effective operators.

A classic example of an EFT is the muon decay

$$\mu \to e\nu_\mu\bar{\nu}_e \, , \tag{5.15}$$

described by the four-Fermi Lagrangian [Fer34]

$$\mathcal{L}_{\mathrm{Fermi}} = \frac{G_{\mathrm{F}}}{\sqrt{2}} \left[\bar{\nu}_e\gamma_\rho(1-\gamma^5)e\right]\left[\bar{\nu}_\mu\gamma^\rho(1-\gamma^5)\mu\right] \, . \tag{5.16}$$

In the SM, $\mathcal{L}_{\mathrm{Fermi}}$ arises in the low-energy limit of the muon four-momentum $q^2 \ll m_W^2$ by "integrating out" the heavy W in the contribution to the scattering amplitude

$$\mathcal{A} \propto g^2 \left[\bar{\nu}_e\gamma_\mu(1-\gamma^5)e\right]\frac{1}{m_W^2 - q^2}\left[\bar{\nu}_\mu\gamma^\mu(1-\gamma^5)\mu\right] \, . \tag{5.17}$$

Thus the W boson is removed as a degree of freedom in the effective theory, but its effect is captured in the effective coupling, the Fermi constant $G_{\mathrm{F}} = \frac{g^2\sqrt{2}}{8m_W^2}$. It is important to note that we have approximated the nonlocal operator appearing in the SM by a local operator in the EFT, resulting in a contact interaction.

Due to the confinement property of QCD, it is impossible to observe the reaction $b \to s\ell^+\ell^-$ directly. But we can observe, for example, the decay $\bar{B}_d(b\bar{d}) \to \bar{K}(s\bar{d})\ell^+\ell^-$, where both the b and the s quark are in a bound state together with a $\bar{d}$ to form mesons. Hence, we need a description of the decay that allows us to separate the different scales. On the one hand, we have the low-energy, nonperturbative physics involving the formation of the meson with the spectator quark at the scale $\Lambda_{\mathrm{QCD}} \approx 0.3\,\mathrm{GeV}$. On the other hand, there are the high-energy, perturbative scales $\mu_b \approx 4\,\mathrm{GeV}$ and $m_W \approx 80\,\mathrm{GeV}$. At a more technical level, EFT takes care of the resummation of large logarithms appearing in perturbation theory with large mass hierarchies. These are the basic reasons for introducing the concept of the $\Delta B = 1$ EFT.

An additional benefit is the independence of the detailed structure of the (fundamental) theory. When looking for NP effects, we consider only extensions of the SM in the sense that every theory has to agree with the SM at tree level, else it would be in contradiction with the experimental facts. In the simplest case, the effect of a particular extension of the SM is to alter only a handful of scalar quantities, the effective coupling constants or *Wilson coefficients* $\mathcal{C}_i$ (see below). It is therefore sufficient to verify that the Wilson coefficients, as dictated by the data, agree with the SM predictions in order to rule out sizable NP contributions. In the more complicated case, an extension of the SM introduces new operators and thus also extra Wilson coefficients arise, to be compared with the vanishing SM predictions.

We define the $\Delta B = 1$ effective theory of $b \to s$ transitions by the effective Hamiltonian [CMM97; BMU00]

$$\mathcal{H}_{\text{eff}} \equiv -\frac{4G_{\text{F}}}{\sqrt{2}} V_{tb} V_{ts}^* \left(\mathcal{H}_{\text{eff}}^{(t)} + \hat{\lambda}_u \mathcal{H}_{\text{eff}}^{(u)} \right) + \text{h.c.}, \qquad \hat{\lambda}_u \equiv V_{ub} V_{us}^* / V_{tb} V_{ts}^*, \tag{5.18}$$

$$\mathcal{H}_{\text{eff}}^{(t)} \equiv \mathcal{C}_1 \mathcal{O}_1^{\text{c}} + \mathcal{C}_2 \mathcal{O}_2^{\text{c}} + \sum_{3 \le i} \mathcal{C}_i \mathcal{O}_i, \qquad \mathcal{H}_{\text{eff}}^{(u)} \equiv \mathcal{C}_1 (\mathcal{O}_1^{\text{c}} - \mathcal{O}_1^{\text{u}}) + \mathcal{C}_2 (\mathcal{O}_2^{\text{c}} - \mathcal{O}_2^{\text{u}}) \,. \tag{5.19}$$

The unitarity triangle (5.12) has been used to split $\mathcal{H}_{\text{eff}}$ into two parts: $\mathcal{H}_{\text{eff}}^{(t)}$ is the dominant contribution for all CP conserving $b \to s$ observables, while all CP violating terms in the SM involve $\mathcal{H}_{\text{eff}}^{(u)}$ that is doubly Cabibbo suppressed; cf. (5.11) and (5.14). We omit the explicit dependence of $\mathcal{C}_i$ on the renormalization scale μ. Throughout, we work at the scale $\mu = 4.2\,\text{GeV}$ and assume $\overline{\text{MS}}$ renormalization.

The dynamics of the light-quark $(q = u, d, s, c, b)$ and leptonic $(\ell = e, \mu, \tau)$ degrees of freedom at the scale of the b quark are described by operators of mass dimension 5 and 6 for the parton transitions $b \to s + (\gamma,\, g,\, \bar{q}q,\, \ell^+\ell^-)$. The SM Wilson coefficients $\mathcal{C}_i$ $(i = 1, \ldots, 10)$ are presently known up to NNLO (and partially NNNLO) in QCD [CMM97; BMU00; MS04; GH05; GHM05; CHM07] and NLO in QED [BGH00; Hub+06; Bob+04; GH01]. This includes the renormalization group evolution (RGE) from the electroweak scale $\mu_W \sim m_W$ down to $\mu_b \sim m_b$, which resums sizable logarithmic corrections to all orders in the QCD coupling α_s [BBL96].

In this work, we want to study NP effects on rare B decays. We choose to work in the SM *operator basis* that is defined in such a way that only the relevant operators — a total of 10 — of dimension 5 and 6 are included and the corresponding 10 Wilson coefficients are real-valued; i.e., the formulation is geared at the $V{-}A$ structure of the SM. This set of assumptions is the simplest (most parsimonious) model-independent extension of the SM, in which each $\mathcal{C}_i$ has a fixed value. The operators due to $b \to s\,\bar{q}q$ transitions are the current-current operators $\mathcal{O}_{1,2}^{u,c}$, the QCD penguin operators for $i = 3, 4, 5, 6$, and the $b \to s$ gluon chromomagnetic dipole operator $i = 8$. Effects of QED penguin operators are neglected since they are small for the decays under consideration. Following the studies of QED corrections to the inclusive decay, we choose the QED coupling α_e at the low scale μ_b, capturing most effects of QED corrections [Bob+04; Hub+06] and removing the main uncertainty due to the choice of the renormalization scheme at LO in QED.

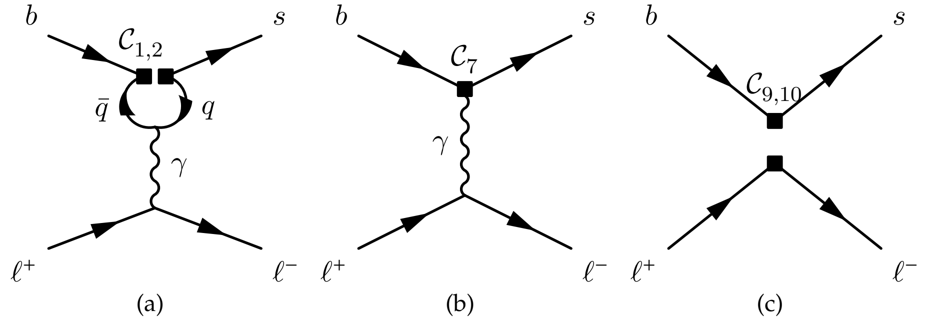


Figure 5.2: Sample FCNC Feynman diagrams of the EFT that describe $b \to s\ell^+\ell^-$ decays. $\mathcal{C}_i$ denotes the coupling strength of the effective vertex.

The electromagnetic dipole operator

$$\mathcal{O}_7 = \frac{e}{(4\pi)^2} m_b \left[\bar{s}\sigma_{\mu\nu}P_R b\right] F^{\mu\nu} \tag{5.20}$$

governs $b \to s\gamma$ transitions. The semileptonic operators

$$\mathcal{O}_9 = \frac{\alpha_e}{4\pi}\left[\bar{s}\gamma_\mu P_L b\right]\left[\bar{\ell}\gamma^\mu \ell\right], \qquad \mathcal{O}_{10} = \frac{\alpha_e}{4\pi}\left[\bar{s}\gamma_\mu P_L b\right]\left[\bar{\ell}\gamma^\mu\gamma_5 \ell\right] \tag{5.21}$$

govern $b \to s\,\ell^+\ell^-$ transitions, cf. Fig. 5.2(c), in combination with less important contributions from $\mathcal{O}_7$ (Fig. 5.2(b)) and from $\mathcal{O}_{1,2}$ (Fig. 5.2(a)).

Beyond the SM, the effects due to new heavy degrees of freedom can be included systematically as additional contributions to the short-distance couplings $\mathcal{C}_i, i = 1 \ldots 10$, possibly giving rise to operators beyond the SM with a different chiral nature or additional light degrees of freedom. Another possible NP effect is extra CP violation due to nonzero imaginary parts of the Wilson coefficients. An important class of NP operators is the chirality-flipped operator basis $\mathcal{O}_i{}'$; e.g.,

$$\mathcal{O}'_{10} = \frac{\alpha_e}{4\pi}\left[\bar{s}\gamma_\mu P_R b\right]\left[\bar{\ell}\gamma^\mu\gamma_5 \ell\right] . \tag{5.22}$$

But in this work, we do not consider $\mathcal{O}_i{}'$ further in accordance with our goal to consider only the *simplest* extension of the SM. The full set of operators is defined in [CMM97, $\mathcal{O}_1 - \mathcal{O}_6$] as well as [BMU00, $\mathcal{O}_7 - \mathcal{O}_{10}$], and summarized in [Dyk12, Section 2.2].

5.3 Nonperturbative effects

When searching for NP in $b \to s$ transitions at the parton level, we need to consider the observable reactions involving B decays. In the particle data group nomenclature, $\bar{B}_q$ is defined as a bound state of one b and one $\bar{q}$ valence quark, $\bar{B}_q \equiv (b\bar{q})$. In this work, we usually consider CP averages, thus B_q represents an admixture of both $(b\bar{q})$ and $(\bar{b}q)$. For further simplification, we mostly omit the subscript q and then restrict to $q = u, d$, but we explicitly display strange B mesons as B_s. The properties of the B and K mesons relevant to this work are listed in Table 5.2. Note that K^* is a vector meson,

	K^+	K^0	K^{*0}	B_u	B_d	B_s
valence quarks	$\bar{s}u$	$\bar{s}d$	$\bar{s}d$	$\bar{b}u$	$\bar{b}d$	$\bar{b}s$
mass [MeV]	493.7	497.6	896.0	5279.3	5279.6	5366.8
J^P	0^-	0^-	1^-	0^-	0^-	0^-

Table 5.2: Properties of selected mesons [Nak+10].

while the other mesons listed are pseudoscalar. For concreteness, we are interested in the exclusive, radiative, and (semi)leptonic, decays

$$B \to K^*\gamma\ , \quad B \to K\ell^+\ell^-\ , \quad B \to K^*\ell^+\ell^-\ , \quad B_s \to \mu^+\mu^-\ , \tag{5.23}$$

with $\ell = e, \mu$. In the previous section, the $\Delta B = 1$ EFT was introduced to separate the short-distance from long-distance behavior. Now we want to briefly sketch the treatment of the long-distance, nonperturbative effects. In naïve factorization, matrix elements of $B \to K\ell^+\ell^-$ decouple into a leptonic and a hadronic part; e.g.

$$\mathcal{A} = \left\langle \bar{\ell}\ell K \middle| \mathcal{C}_9 \mathcal{O}_9 \middle| B \right\rangle = \left\langle \bar{\ell}\ell K \middle| \mathcal{C}_9 \left[\bar{s}\gamma_\mu P_L b\right]\left[\bar{\ell}\gamma^\mu \ell\right] \middle| B \right\rangle \tag{5.24}$$

$$= \mathcal{C}_9 \left\langle \bar{\ell}\ell \middle| \bar{\ell}\gamma^\mu\ell \middle| 0 \right\rangle \left\langle K \middle| \bar{s}\gamma_\mu P_L b \middle| B \right\rangle\ . \tag{5.25}$$

Naïve factorization is justified for interactions described by $\mathcal{O}_9$ and $\mathcal{O}_{10}$ as there are no gluon exchanges between the parts at leading order. Incorporating long-distance effects due to quark loops or gluon exchanges with the spectator quark requires separate approaches depending on the kinematic region. For dilepton invariant mass squared q^2 (see below (5.28)) between $1\,\mathrm{GeV}^2$ and $6\,\mathrm{GeV}^2$, we use *QCD factorization* (QCDF) [BFS01; BFS05]. The charmonium resonances J/ψ and ψ' (cf. the quark loop diagram Fig. 5.2(a)) are dominant in the intermediate region, posing a major problem to the comparison of theory and experiment. Even as we ignore that region, it is important to keep in mind that the tails of the resonances can contribute as much as 20 % even below $6\,\mathrm{GeV}^2$ [Kho+10]. For $q^2 \gtrsim 14\,\mathrm{GeV}^2$, an operator production expansion (OPE) [WZ72; GP04; BBF11] approach that again coincides with naïve factorization at leading order is used. The two regions are displayed in Fig. 5.3.

Contributions to $B \to K^{(*)}\ell^+\ell^-$ from intermediate quark loops — charm loops in particular — (see Fig. 5.2(b)) at leading order in the strong coupling do not require any new hadronix matrix elements compared to those arising from $\mathcal{C}_7, \mathcal{C}_9$, and $\mathcal{C}_{10}$. Hence these long-distance contributions can be absorbed in the *effective* Wilson coefficients

$$\mathcal{C}_7^{\mathrm{eff}} = \mathcal{C}_7 - \frac{1}{3}\left[\mathcal{C}_3 + \frac{4}{3}\mathcal{C}_4 + 20\,\mathcal{C}_5 + \frac{80}{3}\mathcal{C}_6\right] + \frac{\alpha_s}{4\pi}\left[\left(\mathcal{C}_1 - 6\,\mathcal{C}_2\right)A(q^2) - \mathcal{C}_8 F_8^{(7)}(q^2)\right]\ , \tag{5.26}$$

$$\begin{aligned}\mathcal{C}_9^{\mathrm{eff}} = \mathcal{C}_9 &+ h(0,q^2)\left[\frac{4}{3}\mathcal{C}_1 + \mathcal{C}_2 + \frac{11}{2}\mathcal{C}_3 - \frac{2}{3}\mathcal{C}_4 + 52\,\mathcal{C}_5 - \frac{32}{3}\mathcal{C}_6\right] \\ &- \frac{1}{2}h(m_b,q^2)\left[7\,\mathcal{C}_3 + \frac{4}{3}\mathcal{C}_4 + 76\,\mathcal{C}_5 + \frac{64}{3}\mathcal{C}_6\right] + \frac{4}{3}\left[\mathcal{C}_3 + \frac{16}{3}\mathcal{C}_5 + \frac{16}{9}\mathcal{C}_6\right] \\ &+ \frac{\alpha_s}{4\pi}\left[\mathcal{C}_1\left(B(q^2) + 4\,C(q^2)\right) - 3\,\mathcal{C}_2\left(2\,B(q^2) - C(q^2)\right) - \mathcal{C}_8 F_8^{(9)}(q^2)\right] \\ &+ 8\,\frac{m_c^2}{q^2}\left[\left(\frac{4}{9}\mathcal{C}_1 + \frac{1}{3}\mathcal{C}_2\right)(1 + \hat{\lambda}_u) + 2\,\mathcal{C}_3 + 20\,\mathcal{C}_5\right]\ ,\end{aligned} \tag{5.27}$$

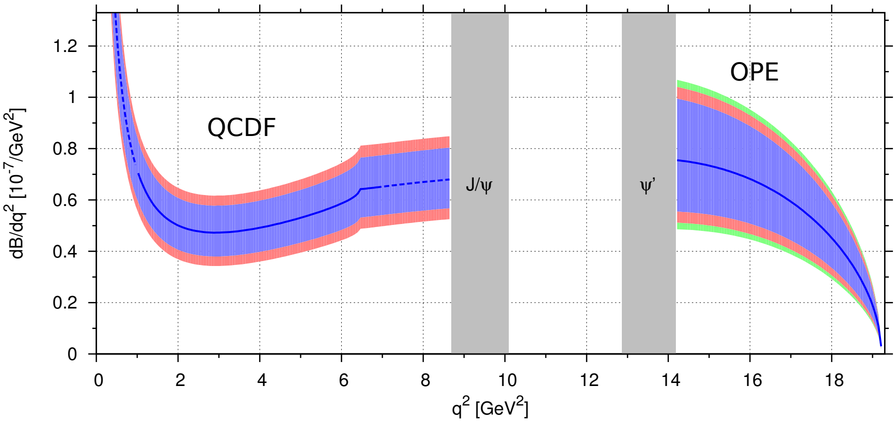


Figure 5.3: The kinematic regions and the different theory approaches to handle nonperturbative effects for the prediction of $\mathcal{B}(B \to K^* \ell^+ \ell^-)$ (solid line). No theory prediction is given for the intermediate region of the charmonium resonances. The blue dashed line indicates extrapolation, the colored bands represent various sources of theory uncertainty (blue: form factors). Reproduced from [BHD10] with the authors' permission.

where we use the results and the nomenclature of [BHD11b].

The leptonic part of $\mathcal{A}$ in (5.24) is calculated in perturbation theory, while the hadronic matrix element is parametrized by *form factors*. Consider $B \to K \ell^+ \ell^-$, or more generally, the overlap of a B and a pseudoscalar meson P. Using the Lorentz transformation properties, the three nonvanishing contributions can be formulated as the scalar, vector, and tensor form factors [BZ05b]

$$\langle P(k)|\,\bar{s}b\,|B(p)\rangle = \frac{m_B^2 - m_P^2}{m_B + m_P} f_0(q^2)\;, \tag{5.28}$$

$$\langle P(k)|\,\bar{s}\gamma_\mu b\,|B(p)\rangle = \left((p+k)_\mu - q_\mu \frac{m_B^2 - m_P^2}{q^2}\right) f_+(q^2) + \frac{m_B^2 - m_P^2}{q^2} q_\mu f_0(q^2)\;, \tag{5.29}$$

$$\langle P(k)|\,\bar{s}\sigma_{\mu\nu} b\,|B(p)\rangle = \frac{\imath}{m_B + m_P}\left[(p+k)_\mu q_\nu - q_\mu (p+k)_\nu\right] f_T(q^2)\;. \tag{5.30}$$

The *transferred four-momentum squared*, $q^2 \equiv (p-k)^2$, pertains to the photon or dilepton, and the three real-valued functions $f_0(q^2)$, $f_+(q^2)$, and $f_T(q^2)$ are the form factors.

Similarly, the five contributions to the overlap of a B meson and a vector meson V can be parametrized in terms of the masses m_B, m_V, the momenta p, k, q, the V polarization vector, and seven form factors $V(q^2)$, $A_{0\text{–}2}(q^2)$, and $T_{1\text{–}3}(q^2)$ [BZ05a]. At large *hadronic recoil*, or equivalently at low q^2, the number of independent form factors reduces to one (P) and two (V) *universal soft form factors* when expanding in $1/m_B$ and relying on large energy effective theory [Cha+99]. In the *low recoil* region, the improved Isgur-Wise relations [GP04; BHD10] — valid up to higher order terms of $\mathcal{O}(\Lambda/Q)\,, Q = m_b, \sqrt{q^2}$ — can be used to relate the dipole form factors $T_i[f_T]$ to the vector and axial form factors V, A_1, and A_2 $[f_+]$. Therefore, the number of independent form factors reduces to four in the case of $B \to V$, and two in $B \to P$.

Nonperturbative QCD effects are taken into account through the form factors. Two competing methods to compute them exist: lattice QCD [Liu+09; AlH+10; Zho+11; Liu+11], and light cone sum rules (LCSR)[BZ05a; BZ05b; Kho+10]. We remark that LCSR is valid only up to $q^2 \lesssim 14\,\text{GeV}^2$, whereas the lattice method works best in the high-q^2 region where the kaon is nearly at rest in the restframe of the decaying B meson. At present, no final results from the lattice are available. While T_1 and T_2 have been determined on the lattice, no such results exist for the vector form factors. Thus, the LCSR predictions are currently extrapolated to larger q^2. The accuracy can be improved by fitting to experimental data; cf. Section 7.2.2 and [HH12]. In general, the form factors are known to a rather low precision of $\mathcal{O}\,(10\,\% - 20\,\%)$, and constitute the major source of theory uncertainty in the prediction of basic observables like the branching ratio $\mathcal{B}$ in which the form factors enter quadratically.

6 Observables and experimental input

The primary goal of this thesis is to search for evidence of new physics in rare B decays using Bayes' theorem (2.4). The essential ingredient in the fit is the likelihood that incorporates the experimental inputs; we list the necessary assumptions in Section 6.1. In the following sections, we take a brief look at where the B decays are actually observed by reviewing two important detector experiments, Belle and LHCb, in Section 6.2. After that, we define the individual observables, grouped according to the decay channel, in Section 6.3. Note that there are two sets of observables: the first contains those that have already been measured and therefore impose constraints on the Wilson coefficients $\mathcal{C}_{7,9,10}$ that we extract in the fit. The second set is made up of observables that are sensitive to the operators of interest and exhibit a reduced hadronic uncertainty, but have not been measured yet; for those we compute improved theory predictions based on the fit output in Section 7.6.

6.1 Basic assumptions

Let us in the following state the assumptions underlying the likelihood and individual observations; as an example, we consider a branching ratio $\mathcal{B}$. Unless indicated otherwise, experimental numbers refer to CP-averaged quantities; i.e., $\mathcal{B}(X \to Y)$ is taken as an abbreviation of

$$\big(\mathcal{B}(X \to Y) + \mathcal{B}(\bar{X} \to \bar{Y})\big)/2 \ . \tag{6.1}$$

We want to stress that we do not use the actual observations — event numbers, event momenta etc. — in our fit. First, it is nearly impossible for us to use them without a precise understanding of the detector. Second, the experiments do not publish the "data" directly, instead they only release a convenient summary of an analysis, typically a maximum-likelihood or maximum-posterior fit in which $\mathcal{B}$ is just one parameter, among many other nuisance parameters modeling the measurement process. Without loss of generality[1], let us assume the posterior $P(\mathcal{B}|D)$ is the 1D function supplied by an experiment, with the understanding that the data are implicit and unknown to us, and only the functional dependence on $\mathcal{B}$ is explicitly given. In order to obtain a contribution to the likelihood from $P(\mathcal{B}|D)$, we "invert" the probability using Bayes' theorem:

$$\begin{aligned} P(\mathcal{B}|D) &\propto \int \mathrm{d}\boldsymbol{\nu}\, \tilde{P}(D|\mathcal{B},\boldsymbol{\nu}) P(\mathcal{B},\boldsymbol{\nu}) \\ &\equiv P(D|\mathcal{B}) \ . \end{aligned} \tag{6.2}$$

As a matter of fact, we define an effective likelihood $P(D|\mathcal{B})$ in (6.2) for use in the fit that implicitly absorbs the effects of the prior $P(\mathcal{B},\boldsymbol{\nu})$ and of the nuisance parameters $\boldsymbol{\nu}$ in the experiment's full likelihood $\tilde{P}(D|\mathcal{B},\boldsymbol{\nu})$. The unknown constant of proportionality is irrelevant for parameter inference and even for model comparison, provided that

[1] A similar argument could be made if we assumed a profile likelihood instead of a posterior.

only Bayes factors are used in which $P(D|\mathcal{B})$ appears in both models. For example, let $P(D|\mathcal{B}) \to cP(D|\mathcal{B})$, then for two models M_1, M_2 with parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ that both predict $\mathcal{B}$ as $\mathcal{B}(\boldsymbol{\theta}_i|M_i)$, the Bayes factor

$$\frac{Z_1}{Z_2} = \frac{c \int \mathrm{d}\boldsymbol{\theta}_1 \, P(D|\mathcal{B}(\boldsymbol{\theta}_1))P(\boldsymbol{\theta}_1)}{c \int \mathrm{d}\boldsymbol{\theta}_2 \, P\,(D|\mathcal{B}(\boldsymbol{\theta}_2))\,P(\boldsymbol{\theta}_2)} \tag{6.3}$$

is unchanged. In contrast, the value of Z_i by itself is arbitrary, thus meaningless. For consistency, we normalize likelihood contributions such that $\int \mathrm{d}\mathcal{B}\, P(D|\mathcal{B}) \equiv 1$, but it really should be $\int \mathrm{d}D\, P(D|\mathcal{B}) \equiv 1$.

In the global fit, we use up to 59 measurements, so let D represent all measurements, not just those used to determine a single observable like $\mathcal{B}$ in the example above. For numerical stability, we work on the log scale; the total log likelihood including the Wilson coefficients $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\nu}$, $\log P(D\,|\,\boldsymbol{\theta}, \boldsymbol{\nu})$ is computed by summing over the individual, independent contributions. The complete list of experimental results used is given in Section 6.3.4 and Tables A.2 to A.4. The majority of results is incorporated as 1D Gaussian distributions, whose variances are obtained by adding statistical and systematic uncertainties in quadrature, $\sigma^2 = \sigma^2_{stat} + \sigma^2_{syst}$. In the case of asymmetric uncertainties, we use a *split Gaussian*,

$$\mathcal{N}(\mathcal{B}|\mu, \sigma_+, \sigma_-) \equiv \begin{cases} \mathcal{N}(\mathcal{B}|\mu, \sigma_+), & \mathcal{B} \geq \mu \\ \mathcal{N}(\mathcal{B}|\mu, \sigma_-), & \mathcal{B} < \mu \end{cases} \tag{6.4}$$

constructed from two half-Gaussian distributions around the central value with variances σ_+ and σ_-. Note that $\mathcal{N}(\cdot|\mu, \sigma_+, \sigma_-)$ is normalized such that 50 % of the probability is on either side of μ, thus if $\sigma_+ \neq \sigma_-$, there is a discontinuity at μ. While this discontinuity — an artifact of the attempt to summarize a probability distribution with just three numbers μ, σ_+, σ_- — is certainly unpleasant, we accept it in order to obtain results that are comparable with the existing literature. We consider it preferable to use the LogGamma distribution (cf. Appendix A.2) to model asymmetric uncertainties in a smooth way, and we do so for asymmetric priors (cf. Appendix A.1).

In summary, the total likelihood is a product of 1D (asymmetric) Gaussians, with the exception of the correlated observables S and C (Section 6.3.1), and the limit on $\mathcal{B}(B_s \to \mu^+\mu^-)$ (Section 6.3.4).

6.2 Experiments

The experiments that contribute to our fit can be grouped into two categories. On the one hand, there are the e^+e^- colliders CESR, PEP-II, and KEKB with the respective detectors CLEO, BaBar, and Belle; on the other hand, the two hadron colliders Tevatron ($p\bar{p}$) and LHC (pp). Of the two general-purpose detectors at Tevatron, CDF and DØ, only CDF has released results on $B \to K^{(*)}\mu^+\mu^-$. At the LHC, there are four detectors: ALICE, ATLAS, CMS, and LHCb. ALICE focuses on a heavy-ion program, and is of no relevance to this work. The two general-purpose detectors ATLAS and CMS, like CDF, feature the ability to accurately identify muons in the final state, and thus are most competitive at detecting the very rare decay $B_s \to \mu^+\mu^-$. The general-purpose detectors collect $B \to K^{(*)}\mu^+\mu^-$ reactions, albeit with a reduced sensitivity as they lack accurate separation of kaons and pions. However, the top priority at CERN is the Higgs search, so ATLAS and CMS have not yet released an analysis of $B \to K^{(*)}\ell^+\ell^-$ as of September 2012.

Last, but certainly not least, LHCb is the detector dedicated to bottom and charm physics at the LHC. Its current analysis of $B \to K^*\mu^+\mu^-$, based on 1 fb^{-1} and 900 candidate events collected in 2011, is already the most accurate in the world (cf. Table A.4). The statistical uncertainties will further decrease with the 2012 data sample, expected to contain an additional 2.2 fb^{-1} and $\mathcal{O}\,(2000)$ events [Hut12]. With regard to $B_s \to \mu^+\mu^-$ searches, ATLAS, CMS, and LHCb are on a similar footing; the current 95 % CL$_S$ limits on $\mathcal{B}(B_s \to \mu^+\mu^-)$ are 22×10^{-9} (2.4 fb^{-1}), 7.2×10^{-9} (5 fb^{-1}), and 4.5×10^{-9} (1 fb^{-1}) respectively [ATL12]. Note that, among the three experiments, LHCb is able to provide the most stringent limit despite the smallest integrated luminosity; more details about how this is accomplished with a special detector setup are given in Section 6.2.2.

In the global fit, we consider these four decays:

$$B \to K^*\gamma\ , \quad B \to K\ell^+\ell^-\ , \quad B \to K^*\ell^+\ell^-\ , \quad B_s \to \mu^+\mu^-\ . \tag{6.5}$$

Note that $\ell = \mu$ is measured by all experiments, but so far, results for $\ell = e$ are available only from BaBar and Belle, as discussed below. Since the experimental accuracy is highest for charged particles in the final state, we only consider the charged kaon $K^\pm$ and the neutral K^{*0} that subsequently decays as $K^{*0} \to K^\pm\pi^\mp$ in the respective decays.

The two channels with the highest statistical impact are $B \to K^*\gamma$ (mostly for $\mathcal{C}_7$), and $B \to K^*\ell^+\ell^-, \ell = \mu$ (mostly for $\mathcal{C}_{9,10}$). The former is best observed at the e^+e^- colliders, and most accurately determined at the two first-generation B factory experiments BaBar and Belle; cf. Table A.2. The latter channel is seen in all experiments; at present, the most accurate measurements are from LHCb. This situation will continue until at least 2016 [Shi11], when first results from Belle II at the second generation B (super flavor) factory SuperKEKB are expected. At the time of writing, it is unclear whether the planned competitor, SuperB, to be located at Frascati, Italy, will receive sufficient funding [Bia+10].

In the following, we present a short overview of the Belle detector representing the first generation B factories, and of LHCb representing a hadron collider detector. With a focus on the decays of interest to us, we highlight the major differences between the two detector concepts. Our review follows [Fuj09] for Belle, and [Ree10] on LHCb.

6.2.1 Belle

The KEKB accelerator and storage ring, located at Tsukuba, Japan, is an e^+e^- collider with about 3 km circumference [Aka+03]. It is the world-record holder in instantaneous luminosity at $2.1 \times 10^{34}\,\text{cm}^2\,\text{s}^{-1}$ and integrated luminosity of 1 ab^{-1}. The electrons and positrons in the beams have energies of 8 GeV and 3.5 GeV respectively, resulting in a center-of-mass energy of $\sqrt{s} = 10.58\,\text{GeV}$ at the interaction point, around which the Belle detector is built. $\sqrt{s}$ is chosen to lie on the $\Upsilon(4S)$ resonance, just above the $B\bar{B}$ production threshold. The branching fraction of $\Upsilon(4S) \to B\bar{B}$ is larger than 96 % [Ber+12]. Because of $2m_B \approx m_{\Upsilon(4S)}$, the quantum-entangled $B\bar{B}$ mesons are produced nearly at rest in the $\Upsilon(4S)$ restframe.

The reason for choosing asymmetric energies of e^+ and e^- is that B mesons receive a Lorentz boost of $\beta\gamma = 0.425$ in the laboratory frame in the direction of the e^- beam axis leading to a mean flight length of 200 µm until the B mesons decay at secondary vertices. This length is large enough to separate the interaction point from the secondary

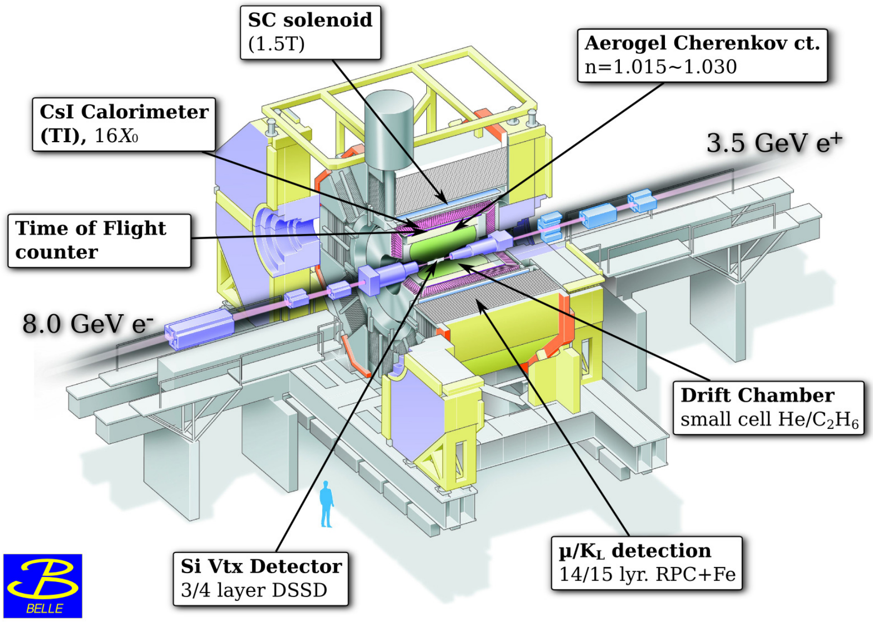


Figure 6.1: The Belle detector at KEKB [Aus12].

vertices, and thus also to separately identify the B and the $\bar{B}$ vertex. The distance Δz between secondary vertices can be translated into the time difference $\Delta t = \Delta z/\beta\gamma$. Δt is crucial to analyzing time-dependent CP asymmetries, one of the main physics concerns of the B factories [Abe+01; Aub+01].

The Belle detector is an asymmetric, large-solid-angle magnetic spectrometer [Aba+02; Nat+06]. It took data from April 2000 until the shutdown in the summer of 2010. In this period, 772×10^6 $B\bar{B}$ pairs were recorded. Belle and the important subdetectors are depicted in Fig. 6.1. The detector is capable of identifying the most commonly arising final state particles in B decays:

$$\text{Charged particles: } K^\pm, \pi^\pm, e^\pm, p^\pm, \mu^\pm$$
$$\text{Neutral particles: } \gamma, K_L^0 \,,$$

where K_L^0 denotes the neutral kaon with the longer lifetime.

In the following, we briefly describe the subdetectors, starting with the innermost part, the *silicon strip vertex detector* (SVD). The first SVD version consists of three concentric cylindrical layers of silicon sensors. This SVD covers a polar angle between 23° and 139° and provide the high resolution needed to separately identify the secondary vertices. After 150×10^6 $B\bar{B}$ pairs, it was replaced with an improved, four-layer SVD that is 1 cm closer to the beam pipe and covers the polar-angle range between 17° and 150°. Overall, the SVD achieves a resolution of Δz between the B vertices better than 100 µm.

The *central drift chamber* (CDC) is a tracking system for charged particles placed around the SVD. It measures particle momenta from the curvature of the helical path that the charged particles follow in the strong 1.5 T field provided by the solenoid magnet; cf. Fig. 6.1. In addition to momenta, the CDC also measures the energy loss $\mathrm{d}E/\mathrm{d}x$ of a particle and thus helps to distinguish electrons, kaons, pions, and protons at momenta below 1 GeV. The particles of interest lose energy by scattering with the gas, a mixture of half helium and half ethane, that permeates the CDC. Location information is obtained from the 8400 drift cells that are mounted inside the CDC. Each cell is made of eight negatively charged wires around a positively charged sense wire.

In order to distinguish charged kaons from charged pions at momenta between 1 and 4 GeV, Bell is equipped with an aerogel Cherenkov counter, a device based on the following effect. Charged particles traveling through a medium, called the *radiator*, at a velocity greater than the speed of light in the medium emit light in a forward cone. The aperture angle of the cone is directly related to the relativistic particle velocity β as

$$\cos\theta = \frac{1}{n_r \beta}\,, \tag{6.6}$$

where n_r is the refractive index of the radiator. In combination with the momentum determination available from the particle's track in the magnetic field of the bending magnet, the mass and thus the identity of the particle is deduced. It is worth noting that this technique does not work for neutral particles. Therefore, the experimental accuracy is significantly better for $B^0 \to K^{*0}\ell^+\ell^-$, where K^{*0} decays into two charged particles, $K^{*0} \to K^\pm\pi^\mp$, as opposed to $B^\pm \to K^{*\pm}\ell^+\ell^-$, where we have $K^{*\pm} \to (K\pi)^\pm$ with only one charged meson in the final state. Experimentally, the most difficult decay is $K^{*0} \to K^0\pi^0$, but this invisible neutral final state is corrected for in the branching ratio. As a consequence, we only use $B^0 \to K^{*0}\ell^+\ell^-$ and $B^\pm \to K^\pm\ell^+\ell^-$ in our global fit.

One of the clever ideas is to use a radiator, the aerogel, with $n_r \gtrsim 1$ chosen such that pions (and electrons) emit Cherenkov light, but the heavier kaons do not because of their lower velocity. A flexible medium, the silica aerogel's refractive index can be chosen by adjusting the manufacturing process; this is exploited to discriminate pions from kaons at different polar angles.

The central drift chamber is supplemented by the *time of flight* (TOF) subdetector to better distinguish kaons from pions and protons at low momenta. A simple on-off device, the TOF measures the time a particle needs to travel from the interaction point to the plastic scintillators of the TOF, where the light is collected in photomultiplier tubes providing a very fast timing resolution of 100 ps.

Identification of electrons and photons is achieved by the *electromagnetic calorimeter* (ECAL) that contains 8736 thallium-doped Cesiumiodide crystals of 30 cm length directed towards the interaction point. Electrons and photons lose energy in the crystals by bremsstrahlung and pair production, while other charged particles transfer energy to the crystal via ionization. The identification of a particle is achieved by comparing the energy estimate in the ECAL and the CDC. The estimates agree more or less for electrons, but differ for other particles. CDF, ATLAS, and CMS have a reduced ability to separate pions from electrons, therefore consider only $\ell = \mu$ in the final state. Similarly, those three experiments do not consider $B \to K^*\gamma$ with the neutral photon in the final state.

Finally, the outermost shell of the detector is given by alternating layers of 5 cm iron blocks and resistive plate counters. The latter are large parallel electrodes with a gas filling the gap. Interactions between hadrons and iron nuclei lead to showers of particles that can ionize the gas and lead to an avalanche. The K^0_L meson lives long enough to reach and strongly interact with the iron, where it is quickly absorbed. But K^0_L does not leave a matching track in the CDC as opposed to charged hadrons. Muons are relatively easy to detect; because they only interact electromagnetically with the iron and

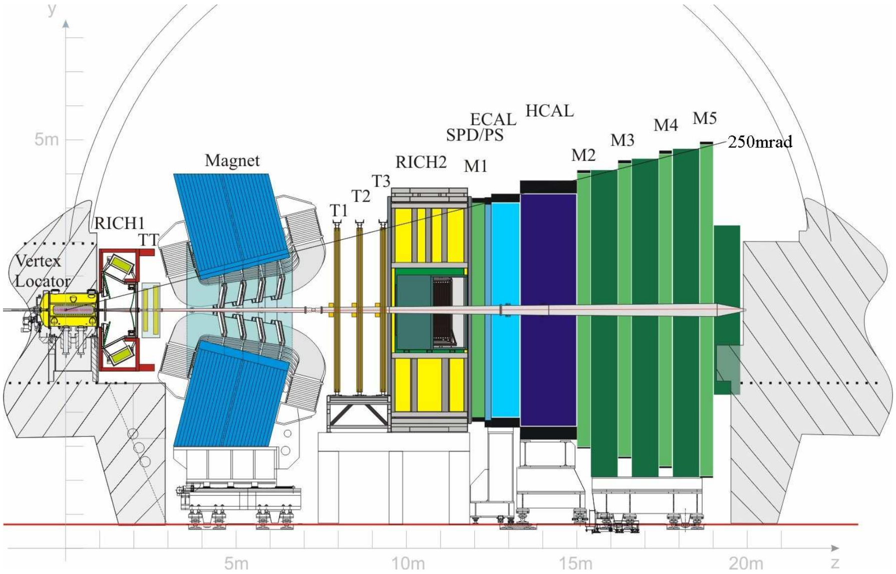


Figure 6.2: A cross section of the LHCb detector. The interaction point is at the left [LHC08].

have a much larger mass than electrons, they have the highest penetration power of all the charged particles expected in the final state. Hence only muons make it through all 15 iron blocks, and their track can be put together with the CDC information.

6.2.2 LHCb

The large hadron collider (LHC) at CERN accelerates two proton beams, revolving in opposite directions in the LEP tunnel of 27 km circumference, and collides them at a center-of-mass energy $\sqrt{s}$ of currently 8 TeV (7 TeV in the 2011 run). Its design luminosity is $10^{34}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$ with $\sqrt{s} = 14\,\mathrm{TeV}$, exceeding the previous record by the Tevatron of $\sqrt{s} = 1.96\,\mathrm{TeV}$ by a factor of 7. The LHCb detector [LHC08] is located at one of the four interaction points, where the beams are focused and brought into collision. At the design energy, the b quark production cross section is at a large value of 500 µb, resulting in the production of 10^{12} $b\bar{b}$ pairs in one year of operation at a reduced luminosity [LHC98]. To achieve the high detection precision, it is desirable to have mostly zero or one $b\bar{b}$ pair per bunch crossing. Therefore, the beams are not focused as strongly as at ATLAS or CMS to level off the luminosity at a value of $2 \times 10^{32}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$. For comparison, the peak luminosity at Belle is larger by a factor of 100. Nevertheless, LHCb has recorded 900 $B \to K^*\ell^+\ell^-$ events during the 2011 run — more than Belle has seen during its entire lifetime (300 events) because of the drastically larger production cross section of 280 µb (LHC at 7 TeV) versus 0.001 µb (KEKB).

The vast majority of $b\bar{b}$ pairs — often hadronizing into $B\bar{B}$ pairs — are produced with momenta in the forward direction from gluon-gluon or quark-quark fusion. Therefore, LHCb is built as a long forward detector to maximize the acceptance and precision at small angles with respect to the beam axis. The interaction point is shifted outside of the cavern hosting the LHCb detector. Thus LHCb covers only one direction in which $B\bar{B}$ pairs are created, but with its length of roughly 20 m, LHCb can achieve great tracking efficiency.

Let us now focus on the most important subdetectors; a cross section of LHCb is shown in Fig. 6.2. $B\bar{B}$ pairs created at LHCb have a large Lorentz boost in the laboratory frame. Combined with the long lifetime of $\approx 1.5\,\mathrm{ps}$ and an average momentum of about 80 GeV, a B meson travels an average distance of 7 mm before decaying. The presence of a secondary vertex, clearly displaced from the primary interaction point, is the number one criterion to identify a B meson. The *vertex locator* is a silicon vertex detector designed to provide high precision tracking near the interaction point. It consists of 21 silicon discs, each split into two overlapping parts. The partition is required to remove the vertex locator from the beam during injection and readjustment of the proton beams. With stable beam conditions, the vertex locator is placed very close to the beam within a distance of only 8 mm, shielded only by a 200 µm thick aluminum foil. Each disc is further divided into smaller segments in a $R - \Phi$ layout, to yield a total of 180 000 read out channels and a precision of roughly 4 µm. The vertex locator is cooled to a temperature of about 270 K to extend its lifetime in the harsh radiation environment.

There are two more silicon trackers, the *tracker turicensis* (TT in Fig. 6.2) and the *inner tracker* (in the center of the tracking stations T1 – T3 in Fig. 6.2). The TT has an active area of $8.4\,\mathrm{m}^2$ to cover the full acceptance region 10 to 250 mrad polar angle, while the inner tracker covers only the small region with the highest occupancy. The outer tracker is made of gas drift tubes filled with Argon and CO_2; the tubes are cheaper and cover a bigger volume, but the large drift time of 50 ns prohibits their use in the inner tracker. The area of the tubes relative to the inner tracker was chosen to have an occupancy of less than 10 % in the drift tubes.

A very important part of LHCb is the *particle identification* provided by the *ring-imaging Cherenkov* (RICH) detector. For our purposes, the most relevant task is to separate kaons from pions, because soft pions from collinear QCD effects can systematically distort the spectra in $B \to K^* \ell^+ \ell^-$.

The RICH system uses the two radiators to increase the identification power. Cherenkov light is created in forward direction and reflected by mirrors onto an array of photon detectors mounted further away from the beam axis. The general-purpose detectors CDF, ATLAS, and CMS lack a good particle identification, in part because of the excessive volume that a RICH system would consume.

Particle energies are determined with the help of the electromagnetic and hadronic calorimeters; the detection mechanism is similar in both. The calorimeters are made of alternating layers of lead and scintillator plastic. The primary particles lose energy in the calorimeter and form a shower of secondary particles, that in turn deposit a fraction of the energy as photons in the scintillator. Those photons are counted by photomultiplier tubes, and with a good understanding of the whole process, the energies of the primary particles are inferred. The calorimeters are segmented, and thus provide additional information for track reconstruction.

Finally, the *muon system* is key to identifying interesting events with B mesons. Muons are used heavily in the level zero trigger; most of the events without muons are discarded right away to reduce the amount of data that needs to be read out, stored, and processed. What makes muons unique is that they have the highest penetrating power of all charged SM particles. The five muon stations (M1 – M5 in Fig. 6.2) are separated by 80 cm thick iron blocks that effectively stop all other particles but muons and neutrinos, the latter escaping undetected. The minimum momentum for a muon to reach M5 is $\approx 6\,\mathrm{GeV}$, but momenta down to 3 GeV are detectable when less station hits are required. Those low-momentum muons are important to accurately filter $B \to K^* \ell^+ \ell^-$ events. Again, the muon stations are segmented to provide tracking information. Multiwire proportional chambers are used except in the most finely granulated section in M1, where gas electron multipliers provide faster readout times needed for the triggering.

6.3 $b \to s$ observables

In the following subsections, we present the individual B decays and describe the observables whose experimental determination enters the global fit. Whenever the likelihood contribution is different from a simple Gaussian, we explain in detail how the information is incorporated. In addition, we discuss observables that have the potential to constrain the Wilson coefficients once they are extrated by the experiments.

6.3.1 $B \to K^* \gamma$ and other radiative decays

For $B \to K^* \gamma$, several observables have been measured, such as the branching ratio $\mathcal{B}$, the time-dependent CP asymmetries S and C, and the isospin asymmetry A_I. Their impact on the scenario of real $\mathcal{C}_{7,7'}$ has been studied in [Des+11; APS12] using the inclusive $\mathcal{B}$ instead of the exclusive one. Here, "inclusive" refers to any final state with a strange meson and a photon, $B \to X_s \gamma$. The measurement of $B_s \to \phi \gamma$ provides similar information and allows a third CP asymmetry H to be studied [MXZ08]. The angular distribution in the decay $B \to K_1(1270)\gamma \to (K\pi\pi)\gamma$ is sensitive to the photon polarization and tests $\mathcal{C}_{7,7'}$. A recent theory study [Bec+12] claims that uncertainties in the range of 20 % for suitably constructed observables are possible at LHCb with 2 fb^{-1}; at present, however, there are no published experimental results for $B \to K_1(1270)\gamma \to (K\pi\pi)\gamma$.

In our analysis, we use $\mathcal{B}$ and the time-dependent CP asymmetries S and C of $B \to K^* \gamma$ with their measurements and correlations at BaBar and Belle compiled in Table A.2, and follow the calculations outlined in [FM03; BFS05]. More details on the numerical input and nuisance parameters can be found in Appendix A.

S and C are interesting due to $B\bar{B}$ mixing, we consider only the case $B \equiv B_d$. With the decay rate Γ and the mass difference $\Delta m_d \approx 3 \times 10^{-10}$ MeV between the heavy and the light mass eigenstate, S and C are extracted simultaneously from [APS12]

$$\frac{\Gamma\left(\bar{B}^0(t) \to \bar{K}^{*0}\gamma\right) - \Gamma\left(B^0(t) \to K^{*0}\gamma\right)}{\Gamma\left(\bar{B}^0(t) \to \bar{K}^{*0}\gamma\right) + \Gamma\left(B^0(t) \to K^{*0}\gamma\right)} = S \sin(\Delta m_d t) - C \cos(\Delta m_d t) \,. \tag{6.7}$$

We include the correlation, as given by the correlation coefficient ρ, into the likelihood as follows. Given the most likely values S^* and C^*, and the total uncertainties σ_S and σ_C, for one experiment, the results are combined into a bivariate Gaussian $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (S^*, C^*)\,, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_S^2 & \rho\sigma_S\sigma_C \\ \rho\sigma_S\sigma_C & \sigma_C^2 \end{pmatrix} \,. \tag{6.8}$$

For the Belle result with asymmetric uncertainties, we set σ_S to the larger of the two uncertainties.

6.3.2 $B \to K\ell^+\ell^-$

In principle, the exclusive decay $B \to K\,\ell^+\ell^-$ offers three observables: the branching ratio $\mathcal{B}(q^2)$, the lepton forward-backward asymmetry $A_{\rm FB}(q^2)$, and the flat term $F_H(q^2)$. The latter two arise in the double-differential decay rate when differentiating with respect to the dilepton invariant mass q^2 and $\cos\theta_\ell$ [BHP07]

$$\frac{1}{\mathrm{d}\Gamma/\mathrm{d}q^2}\frac{\mathrm{d}^2\Gamma}{\mathrm{d}q^2\,\mathrm{d}\!\cos\theta_\ell} = \frac{3}{4}\,(1 - F_H)\sin^2\!\theta_\ell + \frac{1}{2}F_H + A_{\rm FB}\cos\theta_\ell, \tag{6.9}$$

where θ_ℓ is the angle between the 3-momenta of the negatively charged lepton and the $\bar{B}$ meson in the dilepton center of mass system. Two further interesting observables are the rate CP asymmetry $A_{\rm CP}$ and the ratio of decay rates for the ℓ=e and ℓ=μ modes R_K. $A_{\rm FB}$ is nonzero only in the presence of scalar or tensor NP contributions, and F_H is helicity suppressed by $m_\ell/\sqrt{q^2}$ in the scenario under consideration, but is sensitive to scalar and tensor contributions [Bob+01; BHP07]. In view of this, available measurements of $A_{\rm FB}$, F_H, and R_K are not considered, and we include only the $\mathcal{B}$ measurements for one low-q^2 and two high-q^2 bins as listed in Table A.3. Our theory evaluation at low and high q^2 follows [BHP07; Bob+12]. Details concerning numerical input and nuisance parameters are given in Appendix A.

6.3.3 $B \to K^*(\to K\pi)\ell^+\ell^-$

The angular analysis of the 4-body final state $B \to K^*(\to K\pi)\,\ell^+\ell^-$ offers a large set of *angular* observables

$$\langle J_i\rangle\;[q^2_{\rm min}, q^2_{\rm max}] = \int_{q^2_{\rm min}}^{q^2_{\rm max}} \mathrm{d}q^2 J_i(q^2)\,, \qquad\qquad i = 1,\ldots,9\,, \tag{6.10}$$

where the boundaries of the q^2 bin (throughout in units of GeV^2) are not explicitly shown when they are not relevant. The angular observables $\langle J_i\rangle$ are defined in the 3-fold angular distribution

$$\begin{aligned}\frac{32\pi}{9}\frac{\mathrm{d}^3\langle\Gamma\rangle}{\mathrm{d}\!\cos\theta_\ell\,\mathrm{d}\!\cos\theta_K\,\mathrm{d}\phi} \equiv &\\ &\Big[\langle J_{1s}\rangle + \langle J_{2s}\rangle\cos 2\theta_\ell + \langle J_{6s}\rangle\cos\theta_\ell\Big]\sin^2\!\theta_K\\ &+ \Big[\langle J_{1c}\rangle + \langle J_{2c}\rangle\cos 2\theta_\ell + \langle J_{6c}\rangle\cos\theta_\ell\Big]\cos^2\!\theta_K\\ &+ \langle J_3\rangle\sin^2\!\theta_K\sin^2\!\theta_\ell\cos 2\phi + \langle J_4\rangle\sin 2\theta_K\sin 2\theta_\ell\cos\phi + \langle J_5\rangle\sin 2\theta_K\sin\theta_\ell\cos\phi\\ &+ \langle J_7\rangle\sin 2\theta_K\sin\theta_\ell\sin\phi + \langle J_8\rangle\sin 2\theta_K\sin 2\theta_\ell\sin\phi + \langle J_9\rangle\sin^2\!\theta_K\sin^2\!\theta_\ell\sin 2\phi\,.\end{aligned} \tag{6.11}$$

Equation (6.11) accounts for all possible $(\bar{s}\ldots b)(\bar{\ell}\ldots\ell)$ Lorentz structures of chirality-flipped, scalar, pseudoscalar, and tensor operators [KM05; Alt+09; Alo+11]. The explicit dependence of J_i on Wilson coefficients and form factors is presented in [KM05] and [Dyk12, Ch. 3] for the new physics scenario considered in this work. The phase space is parametrized through the dilepton invariant mass squared q^2 and three angles θ_K, θ_ℓ, and ϕ; see also Fig. 6.3. We define θ_K as the angle between the final state kaon and the B meson in the rest frame of K^*, θ_ℓ as the angle between ℓ^- and the B meson in

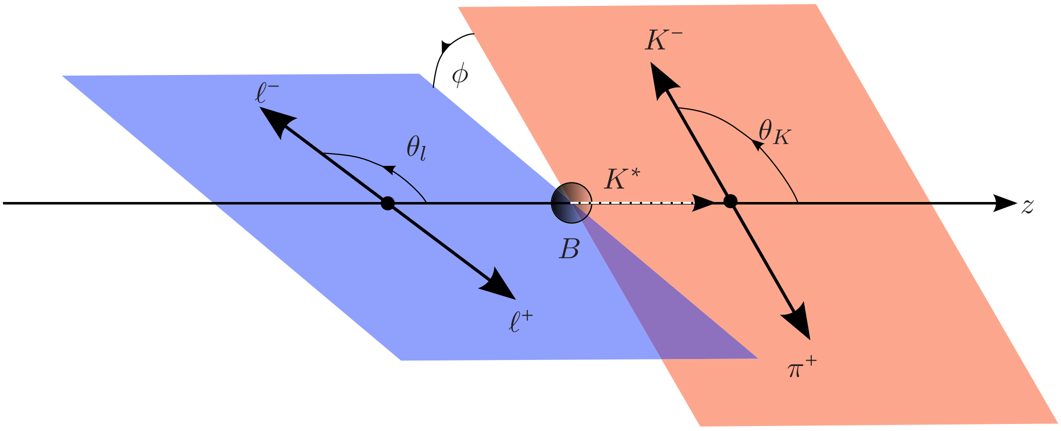


Figure 6.3: The three body decay $B^0 \to K^{*0}\ell^+\ell^-$, with the subsequent decay $K^{*0} \to K^-\pi^+$. The z axis is defined by the direction of K^* in the B rest frame.

the dilepton rest frame exactly as for $B \to K\ell^+\ell^-$, and ϕ describes the angle between the two planes spanned by the lepton momenta and the momenta of the K^* decay products. Our normalization of J_i follows [KM05; Ege+08; Alt+09] and differs by a factor 4/3 from [BHP08; BHD10; BHD11b]. The following simplifications arise in the limit $m_\ell \to 0$ and in the absence of scalar and tensor operators [Alt+09; Alo+11]:

$$J_{1s} = 3\,J_{2s}, \qquad J_{1c} = -J_{2c}, \qquad J_{6c} = 0, \tag{6.12}$$

and a fourth more complicated relation [Ege+10]

$$\begin{aligned} J_{1c} =& 6\frac{(2J_{1s}+3J_3)(4J_4^2+J_7^2)+(2J_{1s}-3J_3)(J_5^2+4J_8^2)}{16J_{1s}^2-9(4J_3^2+J_{6s}^2+4J_9^2)} \\ &-36\frac{J_{6s}(J_4J_5+J_7J_8)+J_9(J_5J_7-4J_4J_8)}{16_{1s}^2-9(4J_3^2+J_{6s}^2+4J_9^2)}\,. \end{aligned} \tag{6.13}$$

Throughout, we assume that the experimental measurements are given for a certain q^2 binning that requires q^2 integration for theory predictions. Consequently, whenever a q^2-dependent observable $X(q^2)$ is defined in a functional form $X(q^2) = f[J_i](q^2)$ in terms of the angular observables, we define the corresponding q^2-integrated quantity as [BHD10]

$$\langle X\rangle = f\left[\langle J_i\rangle\right]. \tag{6.14}$$

It is straightforward to obtain the decay rate and the three single-differential angular distributions from (6.11)

$$\langle\Gamma\rangle = \frac{3}{4}\Big[2\,\langle J_{1s}\rangle + \langle J_{1c}\rangle\Big] - \frac{1}{4}\Big[2\,\langle J_{2s}\rangle + \langle J_{2c}\rangle\Big], \tag{6.15}$$

$$\frac{\mathrm{d}\,\langle\Gamma\rangle}{\mathrm{d}\phi} = \frac{1}{2\pi}\Big[\,\langle\Gamma\rangle + \langle J_3\rangle\cos 2\phi + \langle J_9\rangle\sin 2\phi\Big], \tag{6.16}$$

$$\frac{\mathrm{d}\,\langle\Gamma\rangle}{\mathrm{d}\cos\theta_K} = \frac{3}{8}\Big[\big(3\,\langle J_{1s}\rangle - \langle J_{2s}\rangle\big)\sin^2\theta_K + \big(3\,\langle J_{1c}\rangle - \langle J_{2c}\rangle\big)\cos^2\theta_K\Big], \tag{6.17}$$

$$\frac{\mathrm{d}\,\langle\Gamma\rangle}{\mathrm{d}\cos\theta_\ell} = \frac{3}{8}\Big[2\,\langle J_{1s}\rangle + \langle J_{1c}\rangle + \big(2\,\langle J_{6s}\rangle + \langle J_{6c}\rangle\big)\cos\theta_\ell + \big(2\,\langle J_{2s}\rangle + \langle J_{2c}\rangle\big)\cos 2\theta_\ell\Big]. \tag{6.18}$$

The branching ratio $\langle\mathcal{B}\rangle$, the lepton forward-backward asymmetry $\langle A_{\rm FB}\rangle$, and the longitudinal K^*-polarization fraction $\langle F_L\rangle$

$$\langle\mathcal{B}\rangle = \tau_{B^0}\,\langle\Gamma\rangle\,, \qquad \langle A_{\rm FB}\rangle = \frac{3}{8}\frac{2\,\langle J_{6s}\rangle + \langle J_{6c}\rangle}{\langle\Gamma\rangle}, \qquad \langle F_L\rangle = \frac{3\,\langle J_{1c}\rangle - \langle J_{2c}\rangle}{4\,\langle\Gamma\rangle}, \tag{6.19}$$

have been measured by BaBar [Lee+12; Poi12], Belle [Wei+09], CDF [Aal+11a; Aal+12], and LHCb [Par12]. The angular observable $\left\langle A_T^{(2)}\right\rangle$ [KM05] has been measured by CDF [Aal+12]; and $\langle S_3\rangle$ [Alt+09] has been determined by LHCb [Par12]:

$$\left\langle A_T^{(2)}\right\rangle = \frac{\langle J_3\rangle}{2\,\langle J_{2s}\rangle}, \qquad \langle S_3\rangle = \frac{\langle J_3\rangle}{\langle\Gamma\rangle}. \tag{6.20}$$

The experimental results are summarized in Table A.4. Note that BaBar, Belle, and CDF determine $\langle A_{\rm FB}\rangle$ and $\langle F_L\rangle$ from a combined fit to the single-differential angular distributions

$$\frac{1}{\langle\Gamma\rangle}\frac{\mathrm{d}\langle\Gamma\rangle}{\mathrm{d}\cos\theta_K} = \frac{3}{4}\big[1 - \langle F_L\rangle\,\big]\,\sin^2\theta_K + \frac{3}{2}\,\langle F_L\rangle\cos^2\theta_K, \tag{6.21}$$

$$\frac{1}{\langle\Gamma\rangle}\frac{\mathrm{d}\langle\Gamma\rangle}{\mathrm{d}\cos\theta_\ell} = \frac{3}{4}\,\langle F_L\rangle\sin^2\theta_\ell + \frac{3}{8}\big[1 - \langle F_L\rangle\,\big]\,(1+\cos^2\theta_\ell) + \langle A_{\rm FB}\rangle\cos\theta_\ell \tag{6.22}$$

The observables $\left\langle A_T^{(2)}\right\rangle$ and $\langle A_{\rm im}\rangle = \langle J_9\rangle\,/\,\langle\Gamma\rangle$ are determined from

$$\frac{2\pi}{\langle\Gamma\rangle}\frac{\mathrm{d}\langle\Gamma\rangle}{\mathrm{d}\phi} = 1 + \frac{1}{2}\big[1 - \langle F_L\rangle\,\big]\,\left\langle A_T^{(2)}\right\rangle\cos 2\phi + \langle A_{\rm im}\rangle\sin 2\phi, \tag{6.23}$$

implying $2S_3 = (1 - \langle F_L\rangle)\,\left\langle A_T^{(2)}\right\rangle$. Note that (6.22) and (6.23) are based on the approximation (6.12), which is well justified within our scenario.

The angular observables $\langle J_i\rangle$ and the branching ratio $\langle\mathcal{B}\rangle$ are proportional to the square of hadronic form factors, the main source of theory uncertainty. In normalized combinations of the angular observables, for example $A_{\rm FB}$ and F_L, these uncertainties partially cancel. The most prominent example is the position $q_0^2[A_{\rm FB}]$ of the zero crossing of $A_{\rm FB}$ [Ali+00; BFS01]. It has been determined only recently by LHCb [Par12]; however, we do not include it in the fit due to the large experimental uncertainty. Moreover, note that $q_0^2[A_{\rm FB}]$ is extracted from a fit using mostly the same information as for $\langle A_{\rm FB}\rangle\,[1,6]$, hence including the former in addition to the latter would be double use of the data. A number of suitable combinations of the angular coefficients that reduce the form-factor dependence have been discovered for both low- and high-q^2 regions. At low q^2 [KM05; Ege+08; Ege+10; BS12; Mat+12]

$$\left\langle A_T^{(2)}\right\rangle = \frac{\langle J_3\rangle}{2\,\langle J_{2s}\rangle}, \qquad \left\langle A_T^{(\rm re)}\right\rangle = \frac{\langle J_{6s}\rangle}{4\,\langle J_{2s}\rangle}, \qquad \left\langle A_T^{(\rm im)}\right\rangle = \frac{\langle J_9\rangle}{2\,\langle J_{2s}\rangle}, \tag{6.24}$$

$$\left\langle A_T^{(3)}\right\rangle = \sqrt{\frac{\langle 2\,J_4\rangle^2 + \langle J_7\rangle^2}{-2\,\langle J_{2c}\rangle\,\langle 2\,J_{2s} + J_3\rangle}}\,, \qquad \left\langle A_T^{(4)}\right\rangle = \sqrt{\frac{\langle J_5\rangle^2 + \langle 2\,J_8\rangle^2}{\langle 2\,J_4\rangle^2 + \langle J_7\rangle^2}}\,, \tag{6.25}$$

$$\left\langle A_T^{(5)} \right\rangle = \frac{\sqrt{\langle 4J_{2s}\rangle^2 - \langle J_{6s}\rangle^2 - 4\big(\langle J_3\rangle^2 + \langle J_9\rangle^2\big)}}{8\,\langle J_{2s}\rangle}; \tag{6.26}$$

whereas at high q^2 [BHD10]

$$\left\langle H_T^{(1)} \right\rangle = \frac{\sqrt{2}\,\langle J_4\rangle}{\sqrt{-2\,\langle J_{2c}\rangle\,\langle 2J_{2s} - J_3\rangle}}, \tag{6.27}$$

$$\left\langle H_T^{(2)} \right\rangle = \frac{\langle J_5\rangle}{\sqrt{-2\,\langle J_{2c}\rangle\,\langle 2J_{2s} + J_3\rangle}}, \qquad \left\langle H_T^{(3)} \right\rangle = \frac{\langle J_{6s}\rangle}{2\sqrt{\langle 2\,J_{2s}\rangle^2 - \langle J_3\rangle^2}}. \tag{6.28}$$

For brevity, factors of $\beta_\ell = \sqrt{1 - 4\,m_\ell^2/q^2}$ have been set to unity, since they are negligible in our scenario for the considered range $q^2 \gtrsim 1\,\mathrm{GeV}^2$. Recently, [Mat+12] found that $H_T^{(1)}$ and $H_T^{(2)}$ (P_4 and P_5 in their notation) are also optimized observables at low q^2.

We note that at low q^2, J_3 and J_9 vanish at leading order in QCDF [BHP08], making them ideal probes of chirality-flipped operators $i = 7', 9', 10'$ because leading terms in QCDF are ~ $\mathrm{Re}[\mathcal{C}_i\mathcal{C}_{i'}^*]$ and ~ $\mathrm{Im}[\mathcal{C}_i\mathcal{C}_{i'}^*]$. Only partial results of the subleading corrections exist [BFS05; FM03] and only those of kinematic origin are included in the numerical evaluation. $\left\langle A_T^{(2)} \right\rangle$ and $\langle 2S_3\rangle$ are included in our fit because they might allow us to obtain information on the nuisance parameters used to model yet-unknown subleading contributions (see Appendix A.1.3). J_9 and also $J_{7,8}$ vanish for real Wilson coefficients, and therefore the measurements of $\left\langle A_T^{(\mathrm{im})} \right\rangle$ and $\langle A_{\mathrm{im}}\rangle$ are not of interest in our scenario.

At high q^2, F_L and $A_T^{(2)}$ become short-distance independent [BHD10] and the experimental data allow us to constrain the form-factor-related nuisance parameters; see Section 7.2.2 and Appendix A.1.2. This has been exploited recently [HH12] to extract the q^2 dependence of form factors from data; compared to preliminary lattice results, the authors report overall agreement within the currently sizable uncertainties.

In our predictions, we therefore focus on the yet-unmeasured optimized observables $\left\langle A_T^{(\mathrm{re},3,4,5)} \right\rangle$ at low q^2 and $\left\langle H_T^{(1,2,3)} \right\rangle$ at high q^2.

6.3.4 $B_s \to \mu^+\mu^-$

The rare decay $B_s \to \mu^+\mu^-$ is helicity suppressed in the SM, making it an ideal probe of contributions from scalar and pseudoscalar operators. Its branching ratio depends only on $\mathcal{C}_{10}$ in the scenario under consideration. At leading order in the SM, it is

$$\mathcal{B}(B_s(t=0) \to \mu^+\mu^-) = \frac{G_F^2\,\alpha_e^2\,M_{B_s}^3\,f_{B_s}^2\,\tau_{B_s}}{64\,\pi^3}\left|V_{tb}V_{ts}^*\right|^2\sqrt{1 - \frac{4\,m_\mu^2}{M_{B_s}^2}}\,\frac{4\,m_\mu^2}{M_{B_s}^2}\left|\mathcal{C}_{10}\right|^2 \tag{6.29}$$

and is predicted in the SM to be around 3×10^{-9}. The main uncertainties are due to the decay constant f_{B_s} and the CKM factor $|V_{tb}V_{ts}^*|$. This rare decay is particularly strong in constraining the MSSM parameter $\tan\beta$, as its extra contribution to the (pseudo-) scalar Wilson coefficient is $\propto \tan^3\beta$ [BK00].

In (6.29), the $B_s\bar{B}_s$ mixing has not been taken into account, i.e., the branching ratio refers to time $t = 0$. However, experimentally the time-integrated branching ratio is determined. Both are related in our SM-like scenario as [Bru+12]

$$\mathcal{B}(B_s \to \mu^+\mu^-) = \frac{1}{1-y_s}\mathcal{B}(B_s(t=0) \to \mu^+\mu^-), \qquad y_s = \frac{\Delta\Gamma_s}{2\Gamma_s}. \tag{6.30}$$

Lately, the most precise measurement of the lifetime difference of the two B_s mass eigenstates, $\Delta\Gamma_s$, became available from LHCb [Cla12] and moreover LHCb succeeded to determine the sign of $\Delta\Gamma_s$ [Aai+12a] which turned out to be SM-like. In view of this, we will use the numerical value from LHCb $y_s = 0.088$ [Cla12].

In the last decade, the Tevatron experiments DØ [Aba+10] and CDF [Aal+11b; Cor12] lowered the upper bound on the branching ratio by several orders of magnitude to a value close to 1×10^{-8}; and CDF announced the first direct evidence based on a $2\,\sigma$ fluctuation over the background-only hypothesis [Aal+11b; Cor12]. In 2012, the LHC experiments LHCb, CMS, and ATLAS provided their results based on the complete 2011 run [Aai+12c; Aai+12d; Cha+11; Cha+12b; Aad+12b]. In our analysis we use the most stringent result presented before the 2012 summer conferences, $\mathcal{B}(B_s \to \mu^+\mu^-) < 4.5 \times 10^{-9}$ (3.8×10^{-9}) at 95 % (90 %) CL, obtained by LHCb [Aai+12d]. In the meantime, an improved limit, $\mathcal{B}(B_s \to \mu^+\mu^-) < 4.2 \times 10^{-9}$ (3.7×10^{-9}), based on a combination of the 2012 data of LHCb, CMS, and ATLAS has been published [ATL12]. The experiments report less events than expected within the SM, but overall the agreement is within 1σ.

All of the above limits are extracted using the CL_S method [Rea02]. Now we are facing the question of how to include the search result, or limit, into our likelihood. The CL_S method — a hybrid somewhere in between Bayesian and frequentist territory — is designed with the purpose of either setting a limit or claiming a discovery, but it is not meant to produce a probability distribution about how probable a particular value of $\mathcal{B}$ is. In our opinion, there is no clean way to translate the CL_S curve, much less only two numbers, say $\mathcal{B}_{90}$ and $\mathcal{B}_{95}$, into a useful contribution to the likelihood. However, several schemes of varying sophistication that try to accomplish just that exist in the literature [ATR06; Fla+09; Ree10].

6.3.4.1 Exclusion limit and the Amoroso distribution

We suggest a Bayesian alternative: it is preferable to directly use the posterior on the branching ratio $P(\mathcal{B}\,|\,D)$, computed by a general algorithm for multichannel search experiments [Hei05]. This posterior is almost always produced to compute Bayesian limits for cross checks with CL_S results, but for political reasons, collaborations often decide to include only the CL_S results in a publication. Having obtained $P(\mathcal{B}\,|\,D)$, we can include it directly into the likelihood as in (6.2), thereby including the maximum amount of information. The input numbers — expected signal yields, background yields — that are needed to compute $P(\mathcal{B}\,|\,D)$ are publicly available from LHCb [Aai+12d]; only the correlations of the yields are not published. Our colleagues at LHCb[2] provided the posterior in the form of pairs $(\mathcal{B}_i, F_i)$ with $F_i = i \times 0.01, i = 1, \dots, 98$. Here, F is the posterior cumulative allowing to determine the limit $\mathcal{B}_\alpha$ at credibility level α from

$$F(\mathcal{B}_\alpha|D) \equiv \int_0^{\mathcal{B}_\alpha} \mathrm{d}\mathcal{B}\, P(\mathcal{B}|D) = \alpha\,. \tag{6.31}$$

For convenience, we seek an analytical expression $g(\mathcal{B})$ interpolating the data points. We constrain $g(\cdot)$ by requiring that it vanish for negative branching ratios and that it

[2] Special thanks to Diego Martinez Santos.

yield the same 10 %, 50 %, and 90 % limits as obtained from $F(\cdot|D)$:

$$g(\mathcal{B} \leq 0) = 0 \tag{6.32}$$

$$\int_0^{\mathcal{B}_\alpha} \mathrm{d}\mathcal{B}\, g(\mathcal{B}) = \alpha, \quad \alpha = 0.1, 0.5, 0.9\,. \tag{6.33}$$

We choose $g(\mathcal{B}) = \text{Amoroso}(\mathcal{B}|\mu, \lambda, a, b)$. The Amoroso family [Cro10] is a continuous unimodal four-parameter family of probability distributions that easily accommodates the constraints and provides an accurate approximation. Many well known distributions are direct members or appear as limits of the Amoroso family. Its functional form is

$$\text{Amoroso}(\mathcal{B}|\mu, \lambda, a, b) = \frac{1}{\Gamma(a)}\left|\frac{b}{\lambda}\right|\left(\frac{\mathcal{B}-\mu}{\lambda}\right)^{ab-1}\exp\left[-\left(\frac{\mathcal{B}-\mu}{\lambda}\right)^{b}\right] \tag{6.34}$$

$$\text{for } \mathcal{B},\ \mu,\ \lambda,\ a,\ b\ \in \mathbb{R},\ a > 0,$$
$$\text{support } \mathcal{B} \geq \mu \text{ if } \lambda > 0,\ \mathcal{B} \leq \mu \text{ if } \lambda < 0.$$

We set the location parameter μ to the minimum physical value, $\mu = 0$, and ensure that the scale parameter λ is positive to satisfy (6.32). The values of λ and of the shape parameters a and b are found by numerically solving the set of three equations (6.33). Great care must be taken to select a good initial value for the numerical minimization, else the gradient finding will go astray. We choose initial values similar to the case of the LogGamma distribution, for which details are given in Section A.2. The Amoroso distribution provides an excellent closed-from approximation of the LHCb data; cf. Fig. 6.4. For the input data of

$$\mathcal{B}_{0.1} = 0.56, \quad \mathcal{B}_{0.5} = 2.03, \quad \mathcal{B}_{0.9} = 4.45, \tag{6.35}$$

the relative error is at most 2 %, and the Amoroso parameters are

$$\mu = 0, \quad \lambda = 2.971 \times 10^{-9}, \quad a = 0.824, \quad b = 1.699\,. \tag{6.36}$$

Compared with the CL$_S$ numbers, the Bayesian limits $\mathcal{B}_{90} = 4.45 \times 10^{-9}$, $\mathcal{B}_{95} = 5.32$ come out slightly larger. Finally, it is worth stressing that the posterior mode of the Amoroso PDF is at $\mathcal{B} = 1.27 \times 10^{-9}$ ($\mathcal{B} = 4.2 \times 10^{-9}$ in the SM with the correction proposed in [Bru+12], and not at $\mathcal{B} = 0$, as a naive interpretation of a limit might suggest. Indeed, LHCb reports $\mathcal{B} = \left(0.8^{+1.8}_{-1.3}\right) \times 10^{-9}$ from an unbinned maximum-likelihood fit, in agreement with the mode of $P(\mathcal{B}|D)$. At present, fewer events than expected within the SM are detected, yet it is too early to speak of an anomaly.

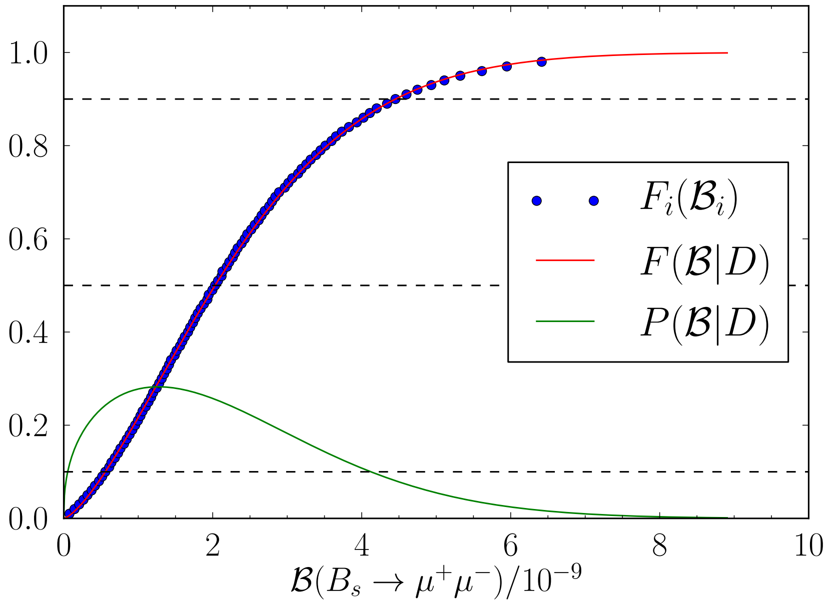


Figure 6.4: The posterior cumulative (dotted blue) from the LHCb search [Aai+12d] for $B_s \to \mu^+\mu^-$ and the interpolation with the cumulative (solid red) of the Amoroso PDF (solid green). The three interpolation fixed points are indicated by horizontal dashed lines.

7 Global fit of rare *B* decays

In our search for new physics in rare B decays, our model M ought to contain minimal assumptions beyond the SM, in particular it should be agnostic to the specifics of the hypothetical underlying physics model; e.g., the MSSM. The effective field theory (EFT) description presented in Section 5.2 is the ideal framework for this purpose. To minimize assumptions about new physics, we assume an EFT with real (instead of complex) Wilson coefficients $\mathcal{C}_i$ and no effective operators beyond the SM basis. In that case, NP is allowed only from heavy, unknown particles in extra contributions to $\mathcal{C}_i$ as

$$\mathcal{C}_i = \mathcal{C}_i^{\mathrm{SM}} + \Delta\mathcal{C}_i^{\mathrm{NP}} \,. \tag{7.1}$$

The motivation is as follows: if there is no discrepancy between the data and the SM, and M fails to give a significantly better explanation of the data, then by Occam's razor it is futile to posit even more involved models than M. Note that this is a purely statistical argument, but it also helps with the physics case. By restricting to $\mathcal{C}_i \in \mathbb{R}$, we explicitly disallow sources of CP violation beyond the CKM mechanism. Reversing the argument, should the need for complex $\mathcal{C}_i$ arise from the data, then we would have a (partial) answer to the CP problem. On a related footing, more involved EFTs have been studied that include additional scalar, vector, and tensor operators [Des+11; BHD11b; APS12; AS12]. If such contributions were found to be present in nature, they would constitute the footprint of a new, more fundamental theory.

7.1 Overview

We analyze the following four $b \to s$ reactions

$$B \to K^*\gamma\,, \quad B \to K\ell^+\ell^-\,, \quad B \to K^*\ell^+\ell^-\,, \quad B_s \to \mu^+\mu^- \tag{7.2}$$

using the $\Delta B = 1$ EFT (cf. Section 5.2) in the SM operator basis (5.18) (5.20) with real Wilson coefficients $\mathcal{C}_i$. Through Bayes' theorem (2.4), we infer the posterior probability of the three parameters of interest, $\boldsymbol{\theta} = (\mathcal{C}_7, \mathcal{C}_9, \mathcal{C}_{10})$, given the data D from experiments. As a convenient diffuse prior, we choose $P(\boldsymbol{\theta}, M) = \text{const}$ in the ranges

$$\mathcal{C}_7 \in [-1, 1], \qquad\qquad \mathcal{C}_{9,10} \in [-10, 10]. \tag{7.3}$$

In order to account for the uncertainty in the theory predictions, we include 28 nuisance parameters $\boldsymbol{\nu}$, six of which due to the CKM parameters (cf. (5.11)) and the masses of the b and c quark are common to all observables. Hadronic uncertainties (cf. Section 5.3) are modeled with two parameters for the $B \to K\ell^+\ell^-$ form factor, one parameter for each of the three form factors appearing in $B \to K^*\gamma, B \to K^*\ell^+\ell^-$ decays , and one parameter for the $B_s \to \mu^+\mu^-$ decay constant. Unknown subleading corrections are separately considered for low and high q^2 due to the different theoretical approaches, QCD factorization (QCDF) at low q^2 and OPE at high q^2 (cf. Section 5.3); they are further subdivided into parameters for $B \to K$ and $B \to K^*$ observables. In total, there are four (K) and twelve (K^*) nuisance parameters for subleading corrections. Full details about each nuisance parameter ν_i and our prior choice $P(\nu_i)$ are given in Appendix A.1. We assume nuisance parameters independent a-priori,

$$P(\boldsymbol{\theta}, \boldsymbol{\nu}, M) = P(\boldsymbol{\theta}, M) P(\boldsymbol{\nu}, M), \qquad P(\boldsymbol{\nu}, M) = \prod_i P(\nu_i, M) \; . \tag{7.4}$$

The likelihood $P(D|\boldsymbol{\theta}, \boldsymbol{\nu}, M)$ is a product of 57 independent terms comprising 59 measurements of 27 observables in the channels (7.2); $B \to K^* \ell^+ \ell^-$ is the dominant contributor with 42 inputs. All observables are defined in Sections 6.3.1 – 6.3.4, their experimental values are summarized in Section 6.3.4.1 and Tables A.2 – A.4. All measurements are included as 1D (asymmetric) Gaussians, except for the bivariate Gaussian describing the two correlated time-dependent CP asymmetries in $B \to K^* \gamma$ (cf. Section 6.3.1), and the Amoroso limit on $\mathcal{B}(B_s \to \mu^+ \mu^-)$ (cf. Section 6.3.4). Bayes' theorem expresses the full 31D posterior as

$$P(\boldsymbol{\theta}, \boldsymbol{\nu}|D, M) = \frac{P(D|\boldsymbol{\theta}, \boldsymbol{\nu}, M) P(\boldsymbol{\theta}, \boldsymbol{\nu}, M)}{Z} \; . \tag{7.5}$$

Samples from the posterior are generated using the PMC algorithm with initialization from MCMC and hierarchical clustering as explained at length in Chapter 4. These draws provide access to all 1D and 2D marginalized distributions as well as the evidence Z. In addition, they are used to perform uncertainty propagation as in (2.16) to compute predictions for observables that have not been measured so far.

We repeat the fit with different settings. One the one hand, by using only subsets of the observables we demonstrate the information coming from, for example, the $B \to K \ell^+ \ell^-$ observables. On the other hand, we run the fit with all observables and two sets of priors to study the prior dependence of our results. Finally, we run the fit within the SM by fixing the Wilson coefficients to SM values and varying only the nuisance parameters for the model comparison with the NP scenario.

7.2 Marginal distributions

7.2.1 Wilson coefficients

Now we present the *main* physics result of this work: the marginal posterior distributions of the Wilson coefficients $\mathcal{C}_{7,9,10}$. The 2D 95 %-credibility regions are shown in Fig. 7.1 when applying the $B \to K^* \gamma$ constraints (Section 6.3.1) in combination with (a) only low- and high-q^2 data from $B \to K \ell^+ \ell^-$ (Section 6.3.2); (b) only low-q^2 data from $B \to K^* \ell^+ \ell^-$ [1]; (c) only high-q^2 data from $B \to K^* \ell^+ \ell^-$ (Section 6.3.3); and finally (d) all the data, including also $B_s \to \mu^+ \mu^-$ (Section 6.3.4).

The most stringent constraints on $\mathcal{C}_{9,10}$ come from the high-q^2 data of $B \to K^* \ell^+ \ell^-$ that should be taken with some caution since the form factors are only available as extrapolations of LCSR results from low q^2. In the near future, we expect more accurate lattice calculations of form factors to close this weak point. Also shown are the SM predictions of $\mathcal{C}_{7,9,10}(\mu = 4.2\,\mathrm{GeV})$ using NNLO evolution [Bob+04].

[1] Here we enlarged the prior ranges of $\mathcal{C}_{7,9,10}$ by a factor of 2.

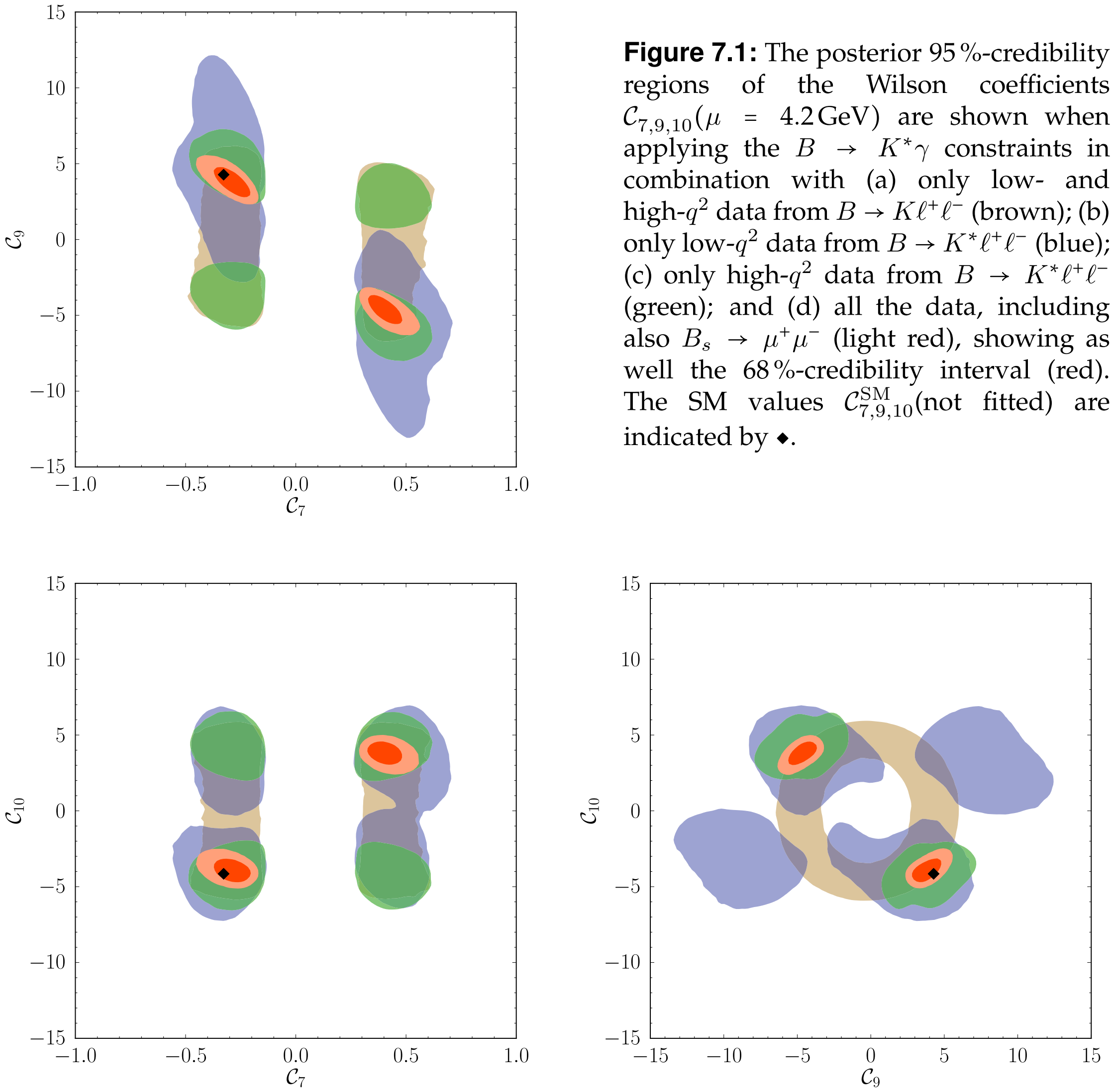


Figure 7.1: The posterior 95 %-credibility regions of the Wilson coefficients $\mathcal{C}_{7,9,10}(\mu = 4.2\,\text{GeV})$ are shown when applying the $B \to K^*\gamma$ constraints in combination with (a) only low- and high-q^2 data from $B \to K\ell^+\ell^-$ (brown); (b) only low-q^2 data from $B \to K^*\ell^+\ell^-$ (blue); (c) only high-q^2 data from $B \to K^*\ell^+\ell^-$ (green); and (d) all the data, including also $B_s \to \mu^+\mu^-$ (light red), showing as well the 68 %-credibility interval (red). The SM values $\mathcal{C}^{\text{SM}}_{7,9,10}$(not fitted) are indicated by ◆.

Using *all* the data, we confirm the findings of previous analyses [APS12; Bob+12] that only two solutions exist at the 95 %-credibility level: the first exhibits the same signs of $\mathcal{C}_{7,9,10}$ as the SM. The second solution corresponds to a first order degeneracy of all observables under a simultaneous sign flip $\mathcal{C}_{7,9,10} \to -\mathcal{C}_{7,9,10}$, which arises as each observable X is a function of products of Wilson coefficients $X = f(\sum_i \sum_{j \geq i} a_{ij} \mathcal{C}_i^{\text{eff}} \mathcal{C}_j^{\text{eff}})$ for some coefficients a_{ij}. Ignoring the small contributions from the operators $\mathcal{O}_1 - \mathcal{O}_6$ to the *effective* Wilson coefficients $\mathcal{C}_i^{\text{eff}}$, the symmetry is exact. There are two additional local maxima that correspond to a sign flip of $\mathcal{C}_7 \to -\mathcal{C}_7$ of the former solutions. In Table 7.1, we list the properties of these four modes, categorized by the signs of $\mathcal{C}_{7,9,10}$. As witnessed by the evidence, Z, the SM-like and sign-flipped solutions essentially make up the whole posterior mass, with ratios of 52 % and 48 %, respectively. The other two solutions are suppressed by many orders of magnitude due to $B \to K^*\ell^+\ell^-$, and thus do not appear at the 95 % level. With the evidence values shown in Table 7.1, the solutions would appear at the 6.3 σ and 7.0 σ level. For the two dominant solutions, the goodness-of-fit results are nearly identical: both p values based on the statistics R_{like}

$\mathrm{sgn}(\mathcal{C}_7, \mathcal{C}_9, \mathcal{C}_{10})$	best-fit point	log(MAP)	R_{like}	$p_{\mathrm{like}}/\%$	R_{pull}	$p_{\mathrm{pull}}/\%$	$\log(Z)$
$(-, +, -)$	$(-0.293, 3.69, -4.19)$	425.22	402.59	60	48.4	75	385.3
$(+, -, +)$	$(0.416, -4.59, 4.05)$	425.08	402.49	60	48.5	75	385.2
$(-, -, +)$	$(-0.393, -3.12, 3.20)$	404.67	387.88	0.9	76.5	4	363.9
$(+, +, -)$	$(0.558, 2.25, -3.24)$	400.91	384.52	0.2	83.1	1	358.9
SM: $(-, +, -)$	$(-0.327, 4.28, -4.15)$	$431.46^{\dagger}$	402.53	70	48.5	83	392.6

Table 7.1: Best-fit point ($\boldsymbol{\nu}$ omitted), log maximum a-posteriori (MAP) value, goodness of fit summary and log evidence for the four local modes (denoted by the signs of $(\mathcal{C}_7, \mathcal{C}_9, \mathcal{C}_{10})$) of the posterior including all experimental constraints. The renormalization scale is fixed to $\mu = 4.2\,\mathrm{GeV}$. For comparison, we include the case with $(\mathcal{C}_7, \mathcal{C}_9, \mathcal{C}_{10})$ fixed at the SM values for which only nuisance parameters are varied (denoted by SM). The nuisance parameters are discarded when counting the degrees of freedom to compute the shown p values based on the statistics R_{like} and R_{pull}. † When comparing the posterior of the SM with the other modes, note that the prior volume of $(\mathcal{C}_7, \mathcal{C}_9, \mathcal{C}_{10})$ is 6.68 in log units.

and R_{pull} (see Section 7.3) are large, indicating a good fit. In contrast, the suppressed solutions do not seem to explain the data well. We note that the MCMC revealed a handful of additional modes with large $6 \lesssim |\mathcal{C}_{9,10}| \lesssim 9$. We do not consider these further because they are suppressed by a factor of roughly $\exp(40)$ compared to the global maximum.

To study the dependence of our fit results on the priors, we use a second set of priors (*wide* priors). We scale the uncertainties of those parameters associated with form factors and unknown subleading contributions in Λ/m_b (Table A.7) by a factor of three and adjust the parameter ranges accordingly. All other priors are kept the same. This choice includes the major sources of theory uncertainty and represents a pessimist's view of (a) the validity of form factor results based on LCSR at low q^2, (b) their extrapolation to high q^2 values, and c) subleading corrections exceeding expectations from power counting. The results of the fit at the low scale $\mu = 4.2\,\mathrm{GeV}$ to all data with these new priors is shown in Fig. 7.2 alongside the corresponding 68 %- and 95 %-credibility regions of Fig. 7.1 for the two solutions in each of the three planes $\mathcal{C}_7 - \mathcal{C}_9$, $\mathcal{C}_7 - \mathcal{C}_{10}$ and $\mathcal{C}_9 - \mathcal{C}_{10}$. Most importantly, the fit is stable and gives comparable results with both sets of priors thanks to the large number of experimental constraints. In all six planes, the area covered by the 68 % region with wide priors is similar to that of the 95 % region with nominal priors. While the two sets of regions are concentric in the $\mathcal{C}_7 - \mathcal{C}_9$ plane, there appears a rather hard cut-off at $|\mathcal{C}_{10}| \approx 5$ in the $\mathcal{C}_{7,9} - \mathcal{C}_{10}$ planes. For completeness, we list the set of smallest intervals and local maxima derived from the 1D marginalized distributions for $\mathcal{C}_{7,9,10}$ for both sets of priors in Table 7.2. Our results for the 95 %-credibility intervals are compatible with those of Ref. [APS12]. More specifically, we find a larger interval for $\mathcal{C}_7$, covering smaller values of $|\mathcal{C}_7|$. This is due to the use of $B \to X_s\gamma$ constraints that are used in Ref. [APS12], but not included in our work. However, with regard to $\mathcal{C}_{9,10}$, our credibility intervals are 10 – 40 % smaller. Compared to Ref. [APS12], we have added the 2012 results by LHCb and BaBar. The question arises if the inclusion of the inclusive decays $B \to X_s\gamma$ and $B \to X_s\,\ell^+\ell^-$ could further shrink the $\mathcal{C}_{9,10}$ credibility intervals. In a future continuation of this work, we expect that stronger constraints on $\mathcal{C}_7$, for example from $B \to X_s\gamma$, will lead to reduced uncertainty also on

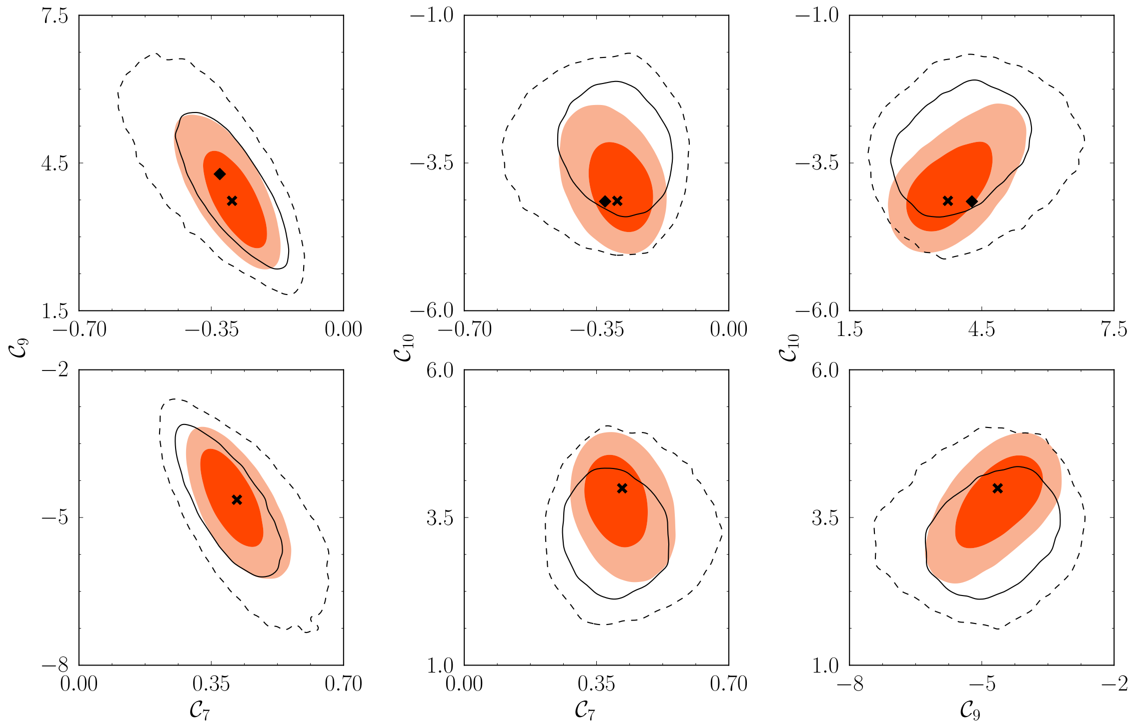


Figure 7.2: The marginalized 2D 68 %- and 95 %-credibility regions of the Wilson coefficients $\mathcal{C}_{7,9,10}$ at $\mu = 4.2\,\text{GeV}$ for the SM-like (top row) and sign-flipped solution (bottom row), arising from nominal ranges as in Fig. 7.1 (red and light red) and wide ranges (solid and dashed contours) of the nuisance parameters. We indicate the values of $\mathcal{C}^{\text{SM}}_{7,9,10}$ in the SM (◆) and at the local maximum of the full posterior (✖) resulting from nominal prior ranges in the respective region.

$\mathcal{C}_{9,10}$ due to the tight connection to $B \to K^*\ell^+\ell^-$ at low q^2.

From the allowed ranges for $\mathcal{C}_{7,9,10}$, we can estimate limits on the scale of generic flavor-changing neutral currents at tree level, described by

$$\mathcal{H}_{\text{eff}} = \sum_{i=7,9,10} \frac{\tilde{\mathcal{O}}_i}{(\Lambda_i^{\text{NP}})^2}\,, \tag{7.6}$$

$$\tilde{\mathcal{O}}_7 = m_b\,[\bar{s}\sigma_{\mu\nu}P_R b]\,F^{\mu\nu}\,, \qquad \tilde{\mathcal{O}}_{9,10} = [\bar{s}\gamma_\mu P_L b][\bar{\ell}\gamma^\mu(1,\gamma_5)\ell]\,. \tag{7.7}$$

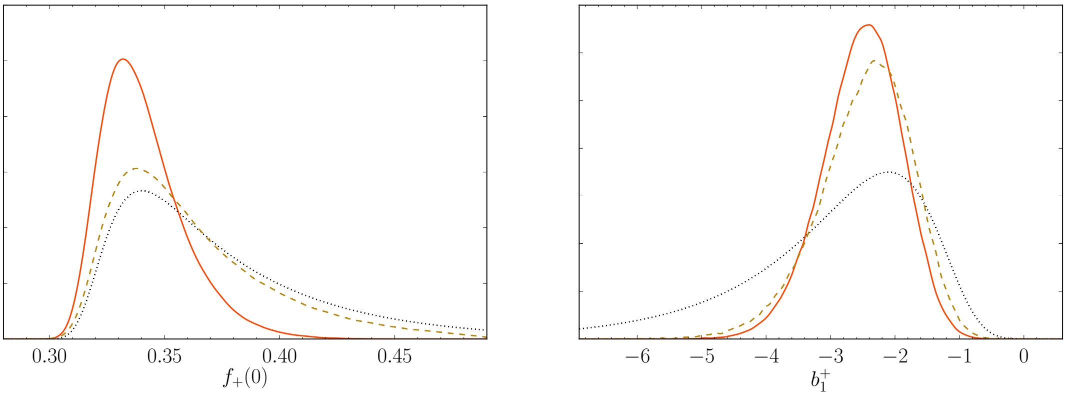
Using $\mathcal{C}_i = \mathcal{C}_i^{\text{SM}} + \Delta\mathcal{C}_i^{\text{NP}}$ and setting $\mathcal{C}_i$ to the boundary values of the 95 % intervals (nominal priors), we extract $\Delta\mathcal{C}_i^{\text{NP}}$. By matching (7.6) with (5.18), (5.20) and (5.21), we extract the minimum scale Λ_i^{NP} for both destructive and constructive interference with the SM; see Table 7.3. The difference between, for example, $\mathcal{O}_9$ and $\tilde{\mathcal{O}}_9$ is that in the latter the coupling is set to one, and the entire suppression is due to the scale $1/\Lambda^{\text{NP}}$, while the former contains SM parameters like $\alpha/4\pi$, V_{ts}, and G_{F}. The resulting scales of $\gtrsim 10\,\text{TeV}$ above which NP “is still allowed” are similar to those found in previous analyses [Bob+12; APS12]. In general, Λ_i^{NP} is about 20 TeV larger in the SM-like solution, but even for the sign-flipped solution with a large deviation from the SM, a direct detection at LHC seems unlikely.

	$\mathcal{C}_7$	$\mathcal{C}_9$	$\mathcal{C}_{10}$
68 %	$[-0.34, -0.23] \cup [0.35, 0.45]$	$[-5.2, -4.0] \cup [3.1, 4.4]$	$[-4.4, -3.4] \cup [3.3, 4.3]$
95 %	$[-0.41, -0.19] \cup [0.31, 0.52]$	$[-5.9, -3.5] \cup [2.6, 5.2]$	$[-4.8, -2.8] \cup [2.7, 4.7]$
modes	$\{-0.28\} \cup \{0.40\}$	$\{-4.56\} \cup \{3.64\}$	$\{-3.92\} \cup \{3.86\}$
68 %	$[-0.39, -0.19] \cup [0.30, 0.48]$	$[-5.6, -3.8] \cup [2.9, 5.1]$	$[-4.0, -2.5] \cup [2.6, 3.9]$
95 %	$[-0.53, -0.13] \cup [0.24, 0.61]$	$[-6.7, -3.1] \cup [2.2, 6.2]$	$[-4.7, -1.9] \cup [2.0, 4.6]$
modes	$\{-0.30\} \cup \{0.38\}$	$\{-4.64\} \cup \{3.84\}$	$\{-3.24\} \cup \{3.30\}$

Table 7.2: The 68 %- and 95 %- credibility intervals and the two local modes of the marginalized 1D posterior distributions of the Wilson coefficients at $\mu = 4.2\,\text{GeV}$, $P(\mathcal{C}_i|D), i = 7, 9, 10$, for nominal (upper) and wide (lower) ranges of nuisance parameters (see Appendix A.1).

	Λ_7^{NP} [TeV]	Λ_9^{NP} [TeV]	Λ_{10}^{NP} [TeV]
SM-like	29, 38	28, 37	30, 44
SM-sign-flipped	12, 13	11, 13	12, 13

Table 7.3: Constraints on the NP scale Λ_i^{NP} ($i = 7, 9, 10$) assuming generic flavor violation at tree level using the 95 %-credibility region from Table 7.2. Two possibilities arise from destructive and constructive interference of the SM with SM-like and SM-sign-flipped solutions.



(a)

(b)

Figure 7.3: Prior (dotted) and posterior distributions of the nuisance parameters $f_+(0)$ (a) and b_1^+ (b), governing the normalization and the q^2 shape of the $B \to K$ form factor $f_+(q^2)$, respectively. We show the posterior using $B \to K\ell^+\ell^-$ data only (dashed) vs all data (solid).

7.2.2 Nuisance parameters

So far, we have discussed the fit results for the Wilson coefficients $\mathcal{C}_{7,9,10}$ that enter most, but not all of the observables. Exceptions are those of $B \to K^*\gamma$, which depends only on $\mathcal{C}_7$, and $B_s \to \mu^+\mu^-$, which depends only on $\mathcal{C}_{10}$. The marginalized distributions in the $\mathcal{C}_9 - \mathcal{C}_{10}$ plane of Fig. 7.1 show that, compared to $B \to K^*$, the fit with $B \to K$ only measurements prefers a smaller value of $|\mathcal{C}_9|^2+|\mathcal{C}_{10}|^2$; the marginal modes (see Fig. 7.8(c)) are near $\mathcal{C}_9 = 0$, $\mathcal{C}_{10} = \pm 5$. Since the $B \to K^*$ constraints dominate the combination, a "tension" arises.

Let us now discuss the role that the nuisance parameters play in the fit. First, we note that the posterior distributions of the common nuisance parameters — those that are not specific to rare $b \to s$ decays, like the CKM parameters and the c and b quark $\overline{\text{MS}}$ masses — do not deviate from their prior distributions given in Table A.5. This is mainly due to the strong prior knowledge from other measurements and the comparatively low precision of both experimental and other, mostly hadronic, theory inputs in the rare $b \to s$ decays.

Second, we consider the remaining hadronic nuisance parameters of form factors and subleading corrections, for which the priors are based mostly on educated guesses rather than precise knowledge. Because $B \to K$ and $B \to K^*$ form factors enter observables at both low and high q^2, they are determined by all the $B \to K\ell^+\ell^-$ and $B \to K^*(\gamma, \ell^+\ell^-)$ observables respectively. In contrast, the parametrization of unknown subleading Λ/m_b corrections is different at low and high q^2 (and naturally in $B \to K$ and $B \to K^*$ decays). Since subleading corrections at high q^2 receive further parametric suppression by either $\mathcal{C}_7/\mathcal{C}_9$ or α_s [GP04; BHD10], the corresponding observables at high q^2 are rather weakly dependent on them. In contrast, at low q^2 large effects are not surprising.

Therefore, we expect a significant update to our knowledge of form factors to accommodate the tension between $B \to K$ and $B \to K^*$ constraints. Any remaining tension should be visible in low-q^2 subleading corrections.

Let us first consider the posterior distributions of the two nuisance parameters $f_+(0)$

and b_1^+ entering the q^2 parametrization of the $B \to K$ form factor $f_+(q^2)$ (see (A.5) and priors in Table A.7 from LCSR results [Kho+10]). The q^2 shape of the form factor is controlled by b_1^+. The low- and high-q^2 data of the $B \to K\ell^+\ell^-$ branching fraction (Table A.3) give rise to a narrower posterior compared to the prior distribution in Fig. 7.3, which does not change much when using only $B \to K\ell^+\ell^-$ data or combining it with $B \to K^*\ell^+\ell^-$. This preference also appears when choosing the wide set of ranges for the prior distributions of the nuisance parameters, demonstrating that the data suppress the tails in the prior of b_1^+. Concerning $f_+(0)$ that corresponds to the normalization of the form factor, we observe a strong preference for low values in the posterior distribution in Fig. 7.3. However, this preference almost disappears when only $B \to K\ell^+\ell^-$ data is used in the fit. This behavior persists even when allowing for wider prior ranges, and is easily understood in terms of the above-mentioned tension.

We also find strong modifications of the posterior with respect to prior distributions for the three scale factors $\zeta_{A_1,A_2,V}$ multiplying the three form factors A_1, A_2, V in $B \to K^*$. The posteriors are shown in Fig. 7.4 along with the common prior distribution. Of the three, A_1 is known most accurately after the fit, while A_2 and V are simultaneously shifted and compressed. Using all constraints, A_1, A_2, and V are shifted towards higher values, but without $B \to K$ constraints, the shift actually points in the opposite direction. Again, the positive shift serves to reduce the tension and allows a good fit to all constraints with values of $\mathcal{C}_{9,10}$ smaller than required by the $B \to K^*$ constraints alone.

Parameters describing subleading phases are mostly unaffected by the fit. All phases come out with a flat distribution, indicating that they could have been omitted from the fit without any consequences. These phases will become important in the future with measurements of CP violating observables.

The largest update to knowledge of subleading parameters occurs for the scale factor of the transversity amplitudes $A_{0,\perp}^L$ (Appendix A.1.3) describing the $B \to K^*$ decays, with a downward shift of about 10 % and a slight reduction of variance. We observe this effect only in the fit with all observables. Neither A_i^R nor $B \to K$ subleading parameters are updated significantly in any of the fits. A_i^R has little effect compared to A_i^L because the observables depend on $A_i^{L,R} \propto \mathcal{C}_9 \mp \mathcal{C}_{10}$, and $\mathcal{C}_9 \approx -\mathcal{C}_{10}$.

There are a number of "optimized" $B \to K^*\ell^+\ell^-$ observables with reduced form factor dependence; cf. (6.24) – (6.28) and the predictions in Section 7.6 However, the opposite is also beneficial. In fact, F_L and $A_T^{(2)}$ are independent of $\mathcal{C}_i$ at high q^2. Nevertheless, we include them in the fit to reduce the theory uncertainty, which in turn helps to improve the posterior knowledge of $\mathcal{C}_i$. This interplay is what makes the global fit powerful. Similarly, $A_T^{(2)}$ depends strongly on subleading corrections at low q^2.

In summary, we do not observe a drastic update of any nuisance parameter, showing that the fit is stable [2]. The uncertainty on the form factors and some subleading corrections is reduced by the data; the most likely values are shifted due to the tension between $B \to K$ and $B \to K^*$ constraints. More theory as well as experimental input is required to reduce the uncertainty on the remaining subleading corrections.

[2] For the suppressed solutions, scale factors for $B \to K^*$ form factors and $A_\perp^L$ shift by $\mathcal{O}(15\,\%)$ and $A_\parallel^L$ even peaks at the left boundary.

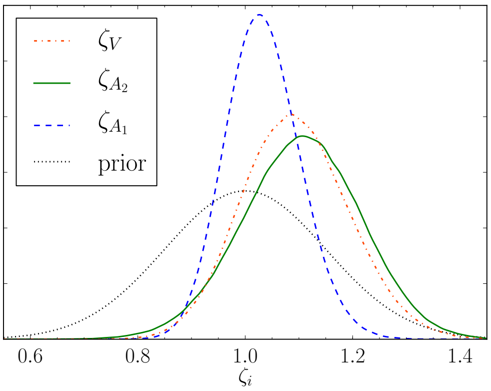


Figure 7.4: Posterior distributions of the fit with all data for the nuisance parameters $\zeta_{A_1,A_2,V}$ serving as scale factors to the corresponding $B \to K^*$ form factors. The common prior is indicated.

7.3 Goodness of fit

To check that the assumed model with three real Wilson coefficients provides a good description of the experimental observations, we determine the goodness of fit. We follow the standard procedure (cf. Appendix C): first we choose a discrepancy variable $R(D|\boldsymbol{\theta}, \boldsymbol{\nu})$ with the parameter values chosen at a local mode of the posterior, then calculate its distribution, and finally determine the p value of the test statistic for the actual data set. For more details on p values and how we interpret them in this work, we refer to [Bea+11]. We make two closely related choices for R, defined as follows.

For each observable X, we compare its theory prediction $X_{\text{pred}}(\boldsymbol{\theta}, \boldsymbol{\nu})$ with the mode of the experimental distribution (central value) of X, denoted by X^*. Next, we compute the frequency f that a value of X less extreme than X_{pred} would be observed. Using the inverse of the Gaussian cumulative distribution function, $\Phi^{-1}(\cdot)$, we define the pull:

$$\delta \equiv \Phi^{-1}\left[\frac{f+1}{2}\right]. \tag{7.8}$$

Note that for a 1D Gaussian, this reduces to the usual $\delta = (X^* - X_{\text{pred}})/\sigma$. In the 1D case, the (Gaussian, Amoroso) distributions yield a signed δ (positive if $X^* > X_{\text{pred}}$, else negative), while for the multivariate Gaussian, δ is positive semidefinite. We define the discrepancy variable R_{pull} as

$$R_{\text{pull}}(D|\boldsymbol{\theta}, \boldsymbol{\nu}) = \sum_i \delta_i^2 \,, \tag{7.9}$$

where i extends over all 57 experimental inputs. In order to pass from the discrepancy variable to a test statistic, we have to evaluate R at *fixed* parameter values; we choose the local modes of the posterior denoted by $\boldsymbol{\theta}^*, \boldsymbol{\nu}^*$. As a cross check to R_{pull}, we also consider R_{like}, defined as the value of the log likelihood, $R_{\text{like}}(D|\boldsymbol{\theta}^*, \boldsymbol{\nu}^*) = \log P(D|\boldsymbol{\theta}^*, \boldsymbol{\nu}^*)$. Its frequency distribution is approximated by generating 10^5 pseudo experiments $D \sim P(D|\boldsymbol{\theta}^*, \boldsymbol{\nu}^*)$. Since we do not have the raw data — events, detector simulations etc. — available, we generate pseudo experiments. Consider the case of a single measurement with Gaussian uncertainties $\mathcal{N}(\mu = X^*, \sigma)$: we fix the theory prediction, shifting the maximum of the Gaussian to $X_{\text{pred}}(\boldsymbol{\theta}^*, \boldsymbol{\nu}^*)$, but keep the uncertainties reported by the experiment. Then we generate $X \sim \mathcal{N}(\mu = X_{\text{pred}}(\boldsymbol{\theta}^*, \boldsymbol{\nu}^*), \sigma)$, and proceed analogously for all observations to sample D. The p value is computed by counting the fraction of experiments with a likelihood value smaller than that for the observed data set and corrected for the number of degrees of freedom; see [Bea+11, Section III.D.5]. Although the generation of pseudo data is far from perfect, we emphasize that, on the one hand, it is fast and, on the other hand, we will not consider the actual value of p too rigorously. Two models with p values of 40 % and 60 % both describe the data well, and that is all the information we want from the p value.

If we used the maximum likelihood parameters and ignored the $B_s \to \mu^+\mu^-$ contribution, both statistics would be equivalent to χ^2 and thus yield the same p value. The parameter values at the global mode of the posterior differ only little from the maximum likelihood values, and the $B_s \to \mu^+\mu^-$ input is negligible as it is only one of 59 inputs. We therefore consider it reasonable to approximate the distribution of R_{pull} by the χ^2-distribution with $\dim D - \dim \boldsymbol{\theta}$ degrees of freedom in order to compute the p value.

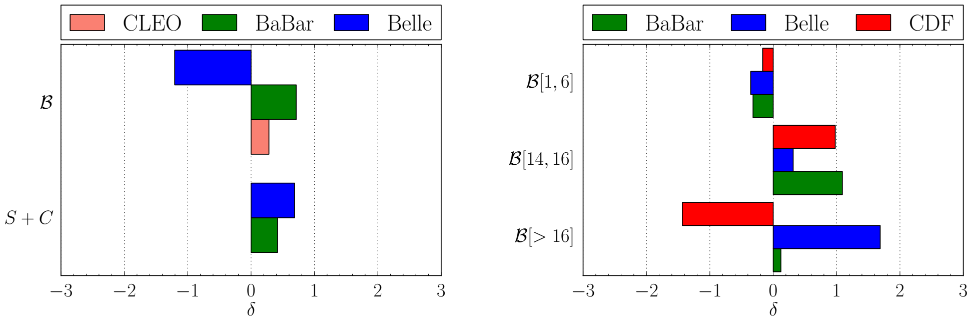
To highlight the excellent quality of the fit, we present the pulls (7.8) for all 57 constraints individually; cf. Fig. 7.5(a) ($B \to K^*\gamma$), Fig. 7.5(b) ($B \to K\ell^+\ell^-$), and Fig. 7.6 ($B \to K^*\ell^+\ell^-$). Finally, the pull for LHCb's result of $B_s \to \mu^+\mu^-$ is -1.1; i.e., its most likely value from the measurement is about 1σ (in terms of the experimental uncertainty) lower than the theory prediction. Here, the theory parameters are chosen at the global maximum of the posterior with SM signature. Fixing $\boldsymbol{\theta}$ at SM values but allowing $\boldsymbol{\nu}$ to vary, we obtain nearly identical plots, and we therefore omit them. The values of R_{like} and R_{pull} are just as good as for the two dominant solutions, but the p values are even larger, as the number of degrees of freedom used in the χ^2-distribution to calculate p differs by three.

We observe the largest pull at +2.5 for the Belle measurement of $\langle\mathcal{B}\rangle\,[16, 19.21]$ for $B \to K^*\ell^+\ell^-$. It is the only pull surpassing 2.0. Fig. 7.6 shows, for example, how the debate about the existence of a zero crossing of A_{FB} at large recoil was settled: the first published measurements by Belle and CDF deviated from the SM prediction, but when taken together with LHCb's recent result that shifts the best-fit point towards the SM, there is good agreement between the SM and the experiments. In fact, LHCb has presented the first direct measurement of the zero crossing based on 1 fb^{-1} at $q_0^2 = \left(4.9^{+1.1}_{-1.3}\right)$ GeV^2 this year [Par12].

In conclusion, the overall goodness of fit as indicated by the p values of R_{pull} in Table 7.1 is very high, and there is not a single very large pull. At the posterior global mode in our scenario M, there are essentially no differences compared with the SM pulls. We conclude that M does not significantly improve over the SM. The data are well described by both fit solutions as well as the SM. p values for R_{like} are uniformly lower, but provide similar support of our findings.

7.4 Model comparison

Since there are several posterior local modes with reasonably high p values, it is necessary to assess which of them is favored by the data; i.e., to perform a model comparison. Suppose the full parameter space is decomposed into disjoint subsets S_i, $i = 1 \dots n$, where S_i contains only a single mode of the posterior. Then we compute the local ev-



(a) (b)

Figure 7.5: Pull values for observables in (a) $B \to K^*\gamma$ and (b) $B \to K\ell^+\ell^-$ calculated at the best-fit point. The pull definition for the correlated observables S and C permits only $\delta \geq 0$; for details see Section 7.3.

idence $Z(S_i)$ by integrating over S_i in (2.5). In fact, $Z(S_i)$ is available as the average weight of all importance samples in S_i, with an accuracy of roughly 5 %.

We also perform the global fit with $\mathcal{C}_{7,9,10}$ fixed to the SM values, varying only the nuisance parameters; see the bottom row in Table 7.1. The prior normalization then changes by $\log(800) = 6.68$ due to omitting $\mathcal{C}_{7,9,10}$ with ranges given in (7.3), and we denote the corresponding evidence by $Z(\text{SM})$. We compute the Bayes factor of the SM vs the SM-like solution by dividing their respective evidences

$$B = \frac{P(D|\text{SM})}{P(D|M)} = \frac{Z(\text{SM})}{Z(M)} = \exp(392.6 - 385.3) \approx 1500 \gg 1. \tag{7.10}$$

Assuming prior odds of one, the posterior odds are given by B,

$$\frac{P(\text{SM}|D)}{P(M|D)} = B\,\frac{P(\text{SM})}{P(M)} = B\,, \tag{7.11}$$

and thus clearly in favor of the simpler model. The effect persists if we cut the prior range of each $\mathcal{C}_i$ in half to exclude all but the SM-like solution, then $Z(S_i) \to Z(S_i)/8$. Similarly, the Bayes factor of SM versus the combination of all local modes is $\approx B/2 \approx 750 \gg 1$.

In conclusion, both the SM (with nuisance parameters allowed to vary) and our extension with real floating $\mathcal{C}_{7,9,10}$ fit the 59 experimental observations of rare B decays well. All the Bayes factor variants are clearly in favor of the SM due its reduced complexity compared to M. In our opinion, prior odds of one do *not* accurately reflect our degree of belief in the SM relative to this particular scenario M considered here; e.g. $P(SM)/P(M) = 1 \times 10^4$ seems more appropriate. Then the posterior odds (2.14) give even stronger support for the SM. In any case, we discouragingly see no evidence of NP in the rare $b \to s$ transitions at the current level of theoretical and experimental precision. However, this may change in the future when optimized $B \to K^*\ell^+\ell^-$ observables with reduced form-factor dependence are measured.

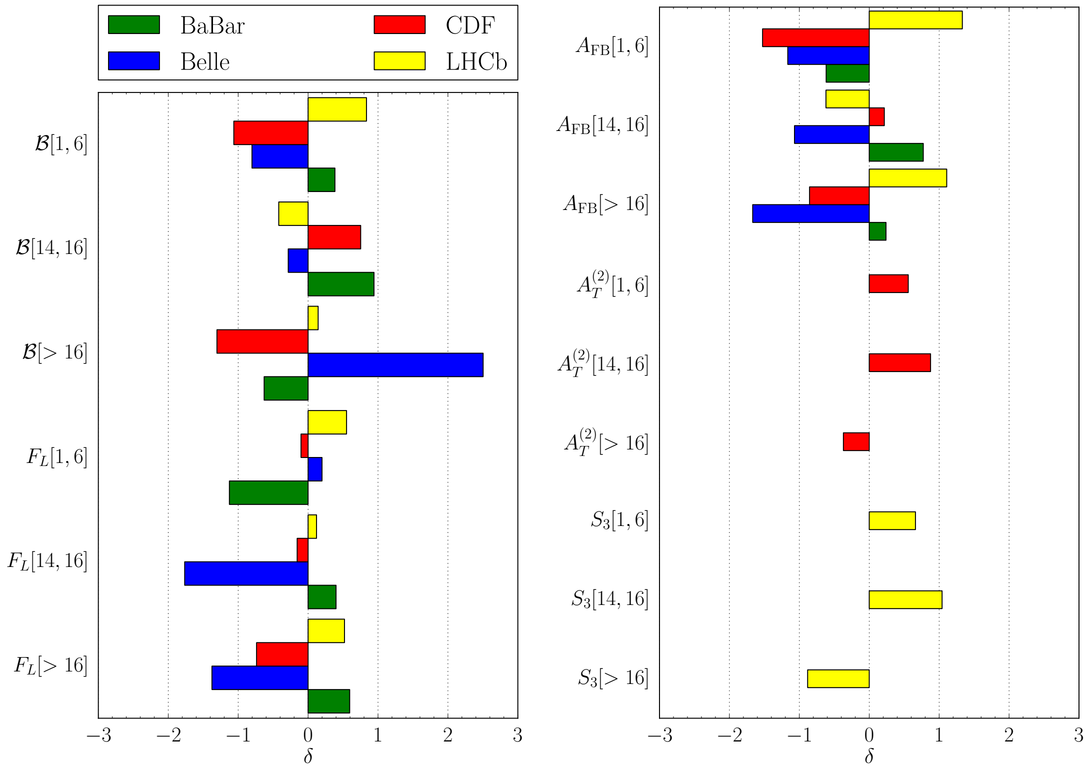


Figure 7.6: Pull values for observables in $B \to K^* \ell^+ \ell^-$ calculated at the best-fit point.

7.5 Sampling performance

Our new combination of MCMC, hierarchical clustering, and PMC detailed in Chapter 4 is the tool we use to obtain samples from the posterior densities in order to derive marginal distributions, to compute the Bayes factors, and to compute predictions through uncertainty propagation. Note that we have to run four independent fits to obtain Fig. 7.1; the likelihood and the number of parameters included in the fit are different for each set of observables. A fifth fit is needed for the set of wide priors to compute Fig. 7.2. Introducing the shorthand notation $B \to K\ell^+\ell^-$ (I), $B \to K^*$@low q^2 (II), $B \to K^*$@high q^2 (III), all data with nominal (IV) and wide (V) priors, and (VI) for the SM, we present the individual fit settings and properties in Table 7.4 for reference. In each fit, the target is far more complex than in the examples that illustrated the clustering in Chapter 4; it is therefore interesting to see how our method fares here. Multiple sources of complexity arise: flat directions are present in the fits (III) – (V) because the uniform priors on the phases of the complex-valued subleading corrections at high q^2 are hardly altered by the fit; most 2D posteriors exhibit significant correlation between parameters. Our results show that all fits converged after $\gtrsim 10$ iterations, independent of the parameter dimensions. Most of the proposal components stay alive throughout the updates, indicating that our construction of the initial proposal is successful.

A particularly beautiful example of the proposal adaptation is shown in Fig. 7.7. The

	d	k	N_{MCMC}	R_c	L	N_g	K_g	N_c	t_{final}	K_{live}	$\mathcal{P}/\%$	ESS /%
(I) $B \to K$	24	40	50 000	2	500	2	45	3000	12	62	32	11
(II) $B \to K^*$@low q^2	18	59	40 000	2	800	4	35	3000	11	126	70	45
(II) $B \to K^*$@high q^2	24	26	40 000	2	800	4	30	3000	14	109	58	32
(IV) all	31	50	60 000	2	1000	4	50	5000	11	95	49	23
(V) wide	31	50	100 000	1.5	1000	5	80	5000	10	78	21	6
(VI) SM	28	30	32000	-	800	1	45	3000	4	43	53	30

Table 7.4: Settings and properties used for the MCMC prerun and the PMC run; symbols as defined in Chapter 4. N_{MCMC} is the length of each chain in the prerun before discarding the burn-in. t_{final} is the number of update steps until convergence; K_{live}, $\mathcal{P}$, and ESS are given for Gaussian components in the final step with 2×10^6 samples and 200 highest-weights cropped.

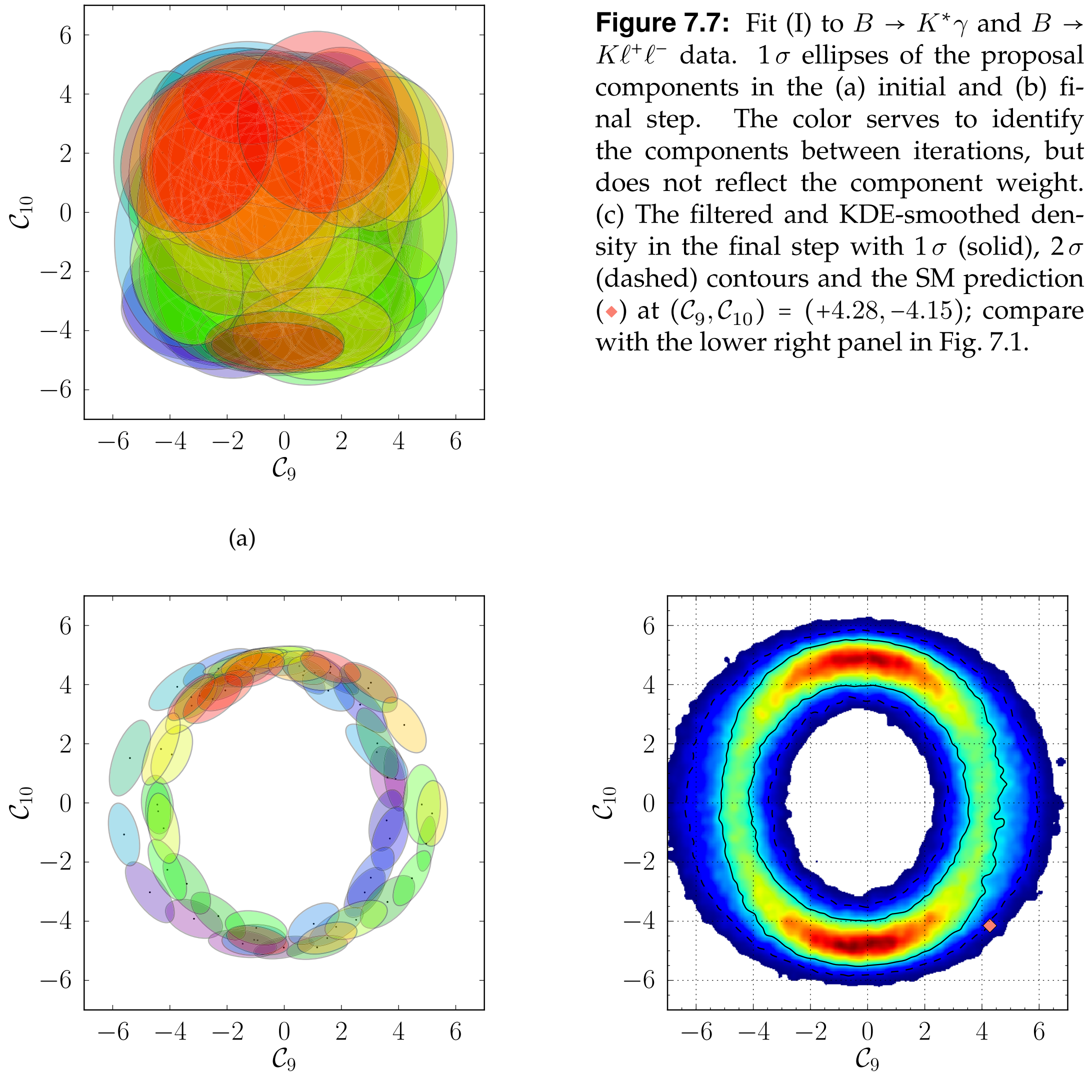
dominant source of knowledge about $\mathcal{C}_9$ and $\mathcal{C}_{10}$ in the 24D $B \to K\ell^+\ell^-$ fit comes from

$$\mathcal{B}(B \to K\ell^+\ell^-) \sim |\mathcal{C}_{79}^{\mathrm{eff}}|^2 + |\mathcal{C}_{10}|^2 \,, \tag{7.12}$$

producing the annulus degeneracy in the $\mathcal{C}_9 - \mathcal{C}_{10}$ plane. The quantity $\mathcal{C}_{79}^{\mathrm{eff}}$ is defined below in (7.14). The 90 components computed by hierarchical clustering overcover the annulus (Fig. 7.7(a)). After the 12 updates, 28 components have died out, the remaining are shrunk and scattered along the annulus, avoiding the low-probability center (Fig. 7.8(b)). The 2D density (Fig. 7.8(c)) shows two banana-shaped local modes around $(\mathcal{C}_9, \mathcal{C}_{10}) = (0, \pm 5)$, but the $1\,\sigma$ and $2\,\sigma$ regions wrap around the entire annulus. Due to this complex shape, both $\mathcal{P}$ and ESS are fairly low, but overall, we consider the adaptation remarkably successful. A single Markov chain traverses only a part of the annulus during the relatively short prerun of length 50 000. But the *ensemble* of chains contains the necessary information about the full annulus.

Common settings among fits (I) – (V) are as follows. In total, we collect 2×10^6 samples in the final step. The marginal densities shown in this chapter are computed from the importance samples with kernel density estimation (KDE) after cropping — those 200 samples with the highest weights are ignored in order to remove outliers, and thereby increasing the ESS. See Section 4.3.3 and Appendix F for more details. The proposal components are of the Gaussian type. Comparing with Student's t with $\nu = 13$ and $\nu = 25$, we observe highest $\mathcal{P}$ and ESS with the Gaussian components. This comes at the expense of a handful severe outliers, but after cropping, the results obtained with the Gaussian proposal have higher ESS, thus lower variance. It is not surprising that Gaussian components give the best results since many of the marginal 1D distributions are approximately Gaussian (see Fig. 7.4), and fat tails present in the priors of $\boldsymbol{\theta}$ and the parameters b_1^+ and $f_+(0)$ (see 7.3) are removed in the posterior by the peaking likelihood.

We verified that our main results — the contours of Fig. 7.1 and Fig. 7.2 — are stable by repeating the three fits with lowest $\mathcal{P}$ — (I), (IV) and (V) — twice with Gaussian and once with Student's t ($\nu = 13$) components. The biggest variations we encountered are the slight "wiggles" in the contours of (V) (see Fig. 7.2) as a consequence of the low ESS = 0.06, the minimum ESS of all fits. In contrast, the contours for fit (IV), also shown in Fig. 7.2, are indistinguishable by eye for repeated runs. The 1D marginal $1(2)-\sigma$ regions and the majority of local mode of Table 7.2 differ only in the least significant



(a)

(b)

(c)

Figure 7.7: Fit (I) to $B \to K^*\gamma$ and $B \to K\ell^+\ell^-$ data. $1\,\sigma$ ellipses of the proposal components in the (a) initial and (b) final step. The color serves to identify the components between iterations, but does not reflect the component weight. (c) The filtered and KDE-smoothed density in the final step with $1\,\sigma$ (solid), $2\,\sigma$ (dashed) contours and the SM prediction (◆) at $(\mathcal{C}_9, \mathcal{C}_{10}) = (+4.28, -4.15)$; compare with the lower right panel in Fig. 7.1.

digit by at most one unit, but a handful of the local modes differs also in next-to-least significant digit by one or two units.

Finally, we scrutinized all marginals by comparing with the MCMC output to ensure all relevant regions of posterior support correctly appear in the PMC output. In the MCMC prerun, we have discarded the first 20 % of the samples for burn-in, and considered the R value for grouping only in $\boldsymbol{\theta}$, but not in $\boldsymbol{\nu}$. By construction, R is a function of a single direction only, and its discriminating power is much larger along a direction $(\boldsymbol{\theta})_i$ compared to $(\boldsymbol{\nu})_i$, because the local modes are well separated only in the $\boldsymbol{\theta}$ direction.

The computations were carried out at the local Max Planck institute for physics (MPP) condor cluster (MCMC prerun only) and at the MPP tier-2 cluster at Rechenzentrum Garching (RZG) (both MCMC and PMC). A single, serial evaluation of the posterior in fit (IV) and (V) with all observables takes about 0.3 s on an Intel i7-2600 operating at 3.4 GHz (MPP condor), and about 0.9 s on an Intel Xeon E5645 (RZG). The major fraction of time is spent on the 2D numerical integration needed for the prediction of binned $B \to K^{(*)} \ell^+ \ell^-$ observables at large recoil; one integration over q^2 and one over the momentum fraction of the s quark in the *light cone distribution amplitude* (LCDA) are required. A total of $\mathcal{O}\left(5 \times 10^6\right)$ posterior evaluations, or at least 17 CPU days, are required for fit (IV) and (V) — the advantage of massive parallelization becomes obvious. We have submitted individual jobs for each of the k chains in the prerun, then merged and split up the chains for the hierarchical clustering that only takes at most 3 min. Finally, the importance sampling has been performed with 200 – 1000 jobs in parallel, but note that the update of the proposal has been done in serial execution, after all jobs computing the importance weights had finished. For the fits in $d = 31$, this update step took a similar amount of time as one of the many jobs computing a small subset of importance weights; we intend to parallelize the updating at the thread level in the future. A massively parallelized update would almost certainly incur no reduction of wallclock time due to the large overhead of starting the jobs and transferring the data; the (binary) output of the final step of the fit with wide priors consumes 560 MB of storage.

7.6 Predictions

As outlined in Section 6.3.3, the angular distribution of $B \to K^*(\to K\pi)\,\ell^+\ell^-$ with its rich set of angular observables gives us the opportunity to form optimized observables that have reduced form factor uncertainties and may exhibit sensitivity to a particular type of new physics. Currently, no measurements of these observables are available. We provide predictions at low and high q^2 within the scenario of the SM operator basis, taking into account the present data. Consequently, future observations outside the predicted ranges would indicate physics beyond the considered scenario. At the technical level, we use the posterior importance samples of the fit with all data and nominal priors to compute samples of an observable X; then we extract the credibility regions from a histogram approximation to the distribution $P(X)$. The procedure is a simple application of uncertainty propagation; cf. Section 2.2.

The predictions of $A_T^{(3,4,5,\mathrm{re})}$ and $H_T^{(1,2)}$ at low q^2 are given in q^2-integrated form for the bin $q^2 \in [1, 6]$ GeV^2 in Table 7.5. In addition, Fig. 7.8 shows the results of the five subbins with a bin width of 1 GeV^2, as used in the first measurement of the lepton A_{FB} of $B \to K^* \ell^+\ell^-$ by LHCb [Par12]. The observables $A_T^{(3,4)}$ have been chosen due to their sensitivity to the chirality-flipped $\mathcal{C}_7'$ [Ege+08]. The large discontinuity of $A_T^{(4)}$ in $q^2 \in [1, 3]$ GeV^2 is caused by the zero crossing of J_4 in its denominator (6.25). The observable $A_T^{(5)}$ is restricted by construction to take values in $[-0.5, 0.5]$ and reaches its maximum value at the zero crossing of the lepton A_{FB} in the bin $q^2 \in [4, 5]$ GeV^2 [Ege+10]. Its shape is sensitive to new physics contributions of the Wilson coefficients. Note that the theory uncertainty is at a minimum when $A_T^{(5)}$ approaches 0.5.

The observable $A_T^{(\mathrm{re})}$ has a peak value of about 1.0 at $q^2 \in [2, 3]$ GeV^2 and has the very same zero crossing as the leptonic A_{FB}. Our results are in qualitative agreement with those of Bećirević and Schneider [BS12], who stress that the deviation of the maximum value from 1.0 and its position are sensitive to new physics. The observables $H_T^{(1,2)}$ were first proposed for the high-q^2 region [BHD10] as long-distance free observables.

In addition, $H_T^{(1)}$ is also short-distance free, with $|H_T^{(1)}(q^2)| = 1$, depending only on the sign of a form factor. Recently it was shown that at low q^2, form factors also cancel in $H_T^{(1,2)}$ [Mat+12]. Each has a zero crossing in the region $q^2 \in [1,3]$ GeV^2 that is the very same as in the CP-averaged normalized observables J_4/Γ and J_5/Γ [Alt+09; BR10]. For $H_T^{(1)}$, one observes the rise towards ≈ 1.0 for rising q^2.

At high q^2, the situation is more restrictive, and within the scenario of the SM operator basis, there are only three optimized observables $H_T^{(1,2,3)}$ [BHD10]. The predictions for three q^2 bins are given in Table 7.6. Besides $|H_T^{(1)}(q^2)| = 1$, we have the additional relation $H_T^{(2)}(q^2) = H_T^{(3)}(q^2)$. Small deviations in the predictions of $\left\langle H_T^{(1,2,3)}\right\rangle$ arise from separate q^2-integration of J_i (see (6.14) and below), such that the equality does not hold exactly. Any large experimental deviation from the prediction $|H_T^{(1)}(q^2)| = 1$ would signal a breakdown of the OPE; cf. Fig. 7.9. The observables $H_T^{(2,3)}(q^2)$ are given by the short-distance ratio [BHD10]

$$H_T^{(2,3)}(q^2) = \frac{2\,\mathrm{Re}\left[\mathcal{C}_{79}^{\mathrm{eff}}(q^2)\,\mathcal{C}_{10}^*\right]}{\left|\mathcal{C}_{79}^{\mathrm{eff}}(q^2)\right|^2 + |\mathcal{C}_{10}|^2} = \cos\left(\varphi_{79}(q^2) - \varphi_{10}\right)\frac{2\,r}{1+r^2}\,, \tag{7.13}$$

with

$$\mathcal{C}_{79}^{\mathrm{eff}}(q^2) = \mathcal{C}_9^{\mathrm{eff}}(q^2) + \kappa\frac{2\,m_b^2}{q^2}\,\mathcal{C}_7^{\mathrm{eff}}(q^2), \qquad\qquad r(q^2) = \frac{|\mathcal{C}_{79}^{\mathrm{eff}}(q^2)|}{|\mathcal{C}_{10}|}\,, \tag{7.14}$$

and $\mathcal{C}_i^{\mathrm{eff}}(q^2)$ and the factor $\kappa = 1 + \mathcal{O}(\alpha_s)$ of the improved Isgur-Wise form factor relation defined in [BHD10]. In the SM, $\mathcal{C}_{10}^{\mathrm{SM}} \approx -4.2$ and therefore its phase is $\varphi_{10} = \pi$. The q^2 dependence of the sum of the effective Wilson coefficients $\mathcal{C}_{79}^{\mathrm{eff}}(q^2)$ is rather weak and its imaginary parts small at NLO in QCD [BHD11b], such that $\varphi_{79}(q^2) \approx 0$; whereas the magnitudes of the Wilson coefficients are $\mathcal{C}_9^{\mathrm{SM}} \approx +4.2$ and $\mathcal{C}_7^{\mathrm{SM}} \approx -0.3$, and lead to $r \approx 1$ and $\cos\left(\varphi_{79}(q^2) - \varphi_{10}\right) \approx -1$. Therefore, $H_T^{(2,3)}$ test roughly the ratio of $|\mathcal{C}_9|/|\mathcal{C}_{10}|$ within our scenario of the SM operator basis and real Wilson coefficients. The results in Table 7.6 show that current data do not allow for deviations from the SM prediction. We remark once again that the prediction of $\left\langle H_T^{(1)}\right\rangle$ is based on the OPE and is expected to be 1 at any particular value of q^2. Therefore, our results just reflect how precisely the form factor and the modeled subleading corrections cancel for the q^2-integrated version when taking into account the update of our knowledge of the nuisance parameters due to the experimental information.

SM predictions have been given before [Alt+09; BHD10; APS12; Bob+12], but our full Bayesian approach provides several improvements with respect to the conventional procedure to estimate theory uncertainties, which we briefly review here. Conventionally, an observable $X(\nu)$ is computed at three values of a single parameter ν: at the central value ν_{cen} and at $(\nu_{\mathrm{cen}})^{+b}_{-a}$. The changes in the prediction of X are then interpreted as the associated uncertainty: $\sigma_{+,-} = |X(\nu_{\mathrm{cen}}) - X(\nu_{\mathrm{cen}}\,{}^{+b}_{-a})|$, and the most probable value of X is assumed to be the central value $X(\nu_{\mathrm{cen}})$. In the presence of several parameters, the respective uncertainties are then combined either linearly or in quadrature into a total uncertainty. In contrast to this so-called min-max approach, we vary all parameters at the same time and thus automatically take correlations into account. Our intervals have a strict probabilistic interpretation as Bayesian credibility intervals, and the procedure automatically takes care of nonlinearities and provides the most probable value.

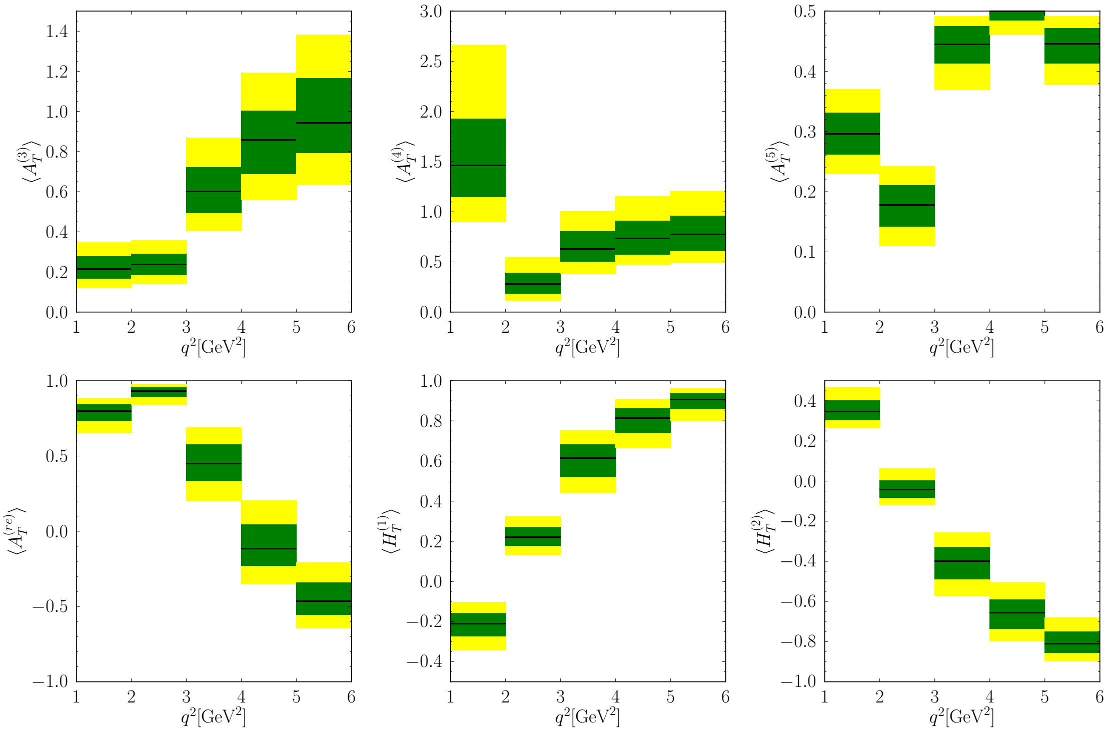


Figure 7.8: Predictions of unmeasured optimized observables at large recoil based on the global fit output. We show the most probable value (solid black line) as well as the smallest 68 % (green) and 95 % (yellow) intervals of the q^2-integrated observables.

q^2-bin	$\langle A_T^{(3)} \rangle$	$\langle A_T^{(4)} \rangle$	$\langle A_T^{(5)} \rangle$
$[1.0,\ 6.0]$	$0.454\ {}^{+0.081}_{-0.086}\ {}^{+0.181}_{-0.158}$	$0.565\ {}^{+0.156}_{-0.121}\ {}^{+0.355}_{-0.234}$	$0.468\ {}^{+0.019}_{-0.025}\ {}^{+0.030}_{-0.056}$
q^2-bin	$\langle A_T^{(\mathrm{re})} \rangle$	$\langle H_T^{(1)} \rangle$	$\langle H_T^{(2)} \rangle$
$[1.0,\ 6.0]$	$0.33\ {}^{+0.14}_{-0.10}\ {}^{+0.25}_{-0.22}$	$0.441\ {}^{+0.055}_{-0.058}\ {}^{+0.105}_{-0.113}$	$-0.271\ {}^{+0.057}_{-0.060}\ {}^{+0.117}_{-0.117}$

Table 7.5: Predictions of unmeasured optimized observables based on global fit output integrated over the *large* recoil region. We list the most probable value and the smallest 68 % and 95 % intervals.

q^2-bin	$\langle H_T^{(1)} \rangle$	$\langle H_T^{(2)} \rangle$	$\langle H_T^{(3)} \rangle$
$[14.18,\ 16]$	$0.99969\ {}^{+0.00009}_{-0.00011}\ {}^{+0.00015}_{-0.00026}$	$-0.9843\ {}^{+0.0023}_{-0.0022}\ {}^{+0.0056}_{-0.0039}$	$-0.9837\ {}^{+0.0022}_{-0.0019}\ {}^{+0.0053}_{-0.0033}$
$[16, 19.21]$	$0.99896\ {}^{+0.00025}_{-0.00032}\ {}^{+0.00044}_{-0.00076}$	$-0.9704\ {}^{+0.0018}_{-0.0019}\ {}^{+0.0042}_{-0.0037}$	$-0.9614\ {}^{+0.0015}_{-0.0012}\ {}^{+0.0037}_{-0.0021}$
$[14.18, 19.21]$	$0.99772\ {}^{+0.00058}_{-0.00078}\ {}^{+0.00105}_{-0.00179}$	$-0.9733\ {}^{+0.0027}_{-0.0023}\ {}^{+0.0057}_{-0.0043}$	$-0.9608\ {}^{+0.0019}_{-0.0015}\ {}^{+0.0045}_{-0.0027}$

Table 7.6: Predictions of unmeasured optimized observables based on global fit output for the two conventional bins and the entire *low* recoil region. We list the most probable value and the smallest 68 % and 95 % intervals.

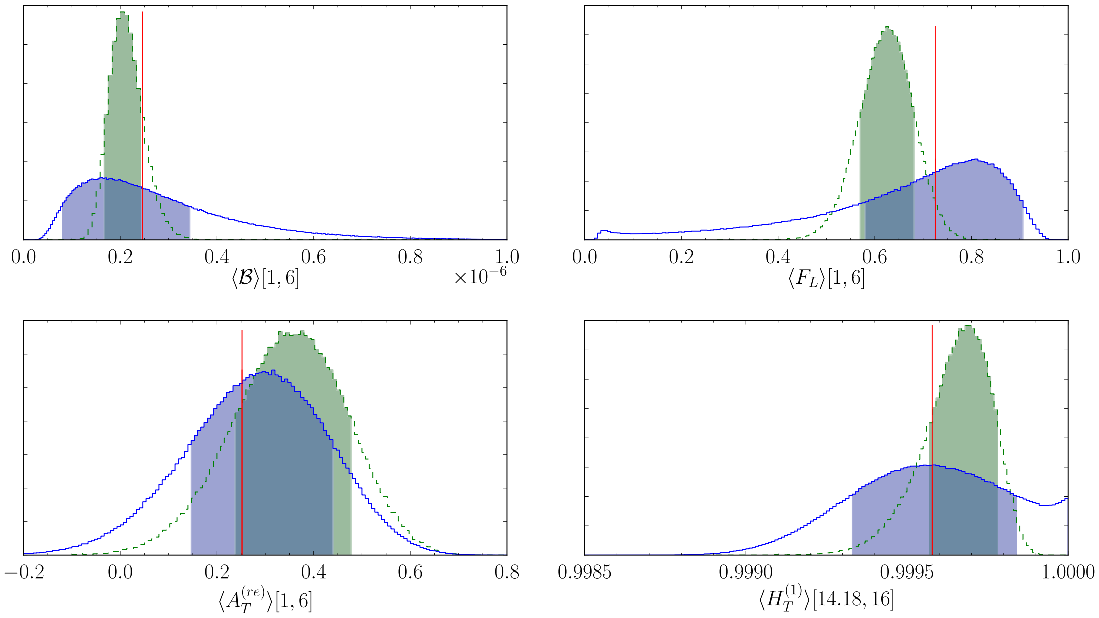


Figure 7.9: Probability distributions of the SM predictions of q^2-integrated observables in the $B \to K^*\ell^+\ell^-$ decay, when varying nuisance parameters within their allowed prior ranges (solid blue). The shaded region is the 68 % interval and the vertical (red) line indicates the prediction when using central values of nuisance parameters. For comparison, we show the predictions based on the posterior of the global fit (dashed green).

As a simple example consider the quadratic dependence of a branching ratio $\mathcal{B}$ on a decay constant or form factor f, $\mathcal{B} \propto f^2$. Assuming a Gaussian prior distribution for f due to MAXENT, $P(\mathcal{B})$ is the asymmetric χ^2-distribution with one degree of freedom. Typical examples of such asymmetry can be seen in in Fig. 7.9 for $\langle\mathcal{B}\rangle\,[1,6]$ and $\langle F_L\rangle\,[1,6]$ (blue, solid) of the decay $B \to K^*\ell^+\ell^-$, where the maximum of the distribution deviates from the vertical (red) line that indicates the prediction obtained by fixing the nuisance parameters to their prior modes; i.e., the red line denotes the central values of the min-max approach. This asymmetric behavior is not present in $\left\langle A_T^{(\mathrm{re})}\right\rangle[1,6]$ since the form factors cancel; likewise in $\left\langle H_T^{(1)}\right\rangle[14.18,\,16]$. We list the modes and 68 % intervals for a number of observables in Appendix B in Table B.1 and Table B.2, but stress that the uncertainty of an observable X is described by the probability distribution $P(X)$. In the simplest Gaussian case, $P(X)$ can be described by the mode and the 68 % interval (for example $\left\langle A_T^{(\mathrm{re})}\right\rangle[1,6]$), but in general, $P(X)$ contains more information. As an example, consider the boundary mode in the distribution of $P\left(\left\langle H_T^{(1)}\right\rangle[14.18,\,16]\right)$ in Fig. 7.9.

Our SM predictions are in good agreement with the existing literature, in which the min-max approach is used throughout. Given that we use the same codebase of EOS [Dyk+12], our results closely match those of Bobeth, Hiller, and Dyk [BHD10] and Bobeth et al. [Bob+12]. Minor differences in the central value and the magnitude of the (asymmetric) uncertainties are due to the different statistical approach as described above. Specifically, all central values are contained in the minimal 68 % regions of the Bayesian results. Compared to the results of Altmannshofer, Paradisi, and Straub

[APS12], our modes are contained in their uncertainty intervals and vice versa. But for some observables, the size of the uncertainty region differs drastically. For example, compare our result

$$\langle \mathcal{B}(B \to K^* \ell^+ \ell^-) \rangle \, [1, 6] = 1.64 \, ^{+1.80}_{-0.83} \tag{7.15}$$

to [APS12]

$$\langle \mathcal{B}(B \to K^* \ell^+ \ell^-) \rangle \, [1, 6] = 2.28 \pm 0.63 \, . \tag{7.16}$$

On the other hand, we also observe instances in which our region is much smaller. For example, for F_L at high q^2, which is independent of $\mathcal{C}_i$, but strongly dependent on form factors, we find

$$\langle F_L(B \to K^* \ell^+ \ell^-) \rangle \, [> 16] = 0.35 \, ^{+0.02}_{-0.03} \tag{7.17}$$

and [APS12]

$$\langle F_L(B \to K^* \ell^+ \ell^-) \rangle \, [> 16] = 0.34 \, \pm 0.22 \, . \tag{7.18}$$

These discrepancies are a consequence not only of min-max versus Bayes, but of the different ways hadronic and parametric uncertainties are taken into account; e.g, all seven $B \to K^* \ell^+ \ell^-$ form factors at low q^2 are kept in [APS12], while we use form factor relations to arrive at only two independent form factors (Section 5.3).

Let us finally compare the SM predictions of observables based on the prior information with predictions based on the posterior distribution of the global fit that includes also NP in $\mathcal{C}_{7,9,10}$. Our posterior findings are overlaid on the SM predictions for the examples in Fig. 7.9. Although the Wilson coefficients are (unconstrained) a-priori, in all cases the posterior predictions are narrower than the SM prediction based on prior knowledge only. Apparently, the additional information in the data on both $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ — in particular the form factors — more than compensates the effect of floating $\boldsymbol{\theta}$, which shows the benefit of our statistical approach. The difference is less pronounced for observables with reduced hadronic uncertainty such as $\left\langle A_T^{(\mathrm{re})} \right\rangle [1, 6]$, where both predictions are of similar quality, the global-fit prediction shifted slightly towards larger values, compared to the SM prediction based on prior knowledge alone.

The same situation emerges for the other optimized observables where main uncertainties are due to lacking subleading corrections; compare Table 7.5 and Table B.1 for low-q^2 as well as Table 7.6 and Table B.2 for high q^2. At this stage, the precision can only improve with better prior knowledge of the nuisance parameters. This will help to distinguish NP from the SM with the help of optimized observables in the scenario of the SM operator basis with real Wilson coefficients. However, any experimental observation outside of the predicted range would point strongly to an extended scenario.

7.7 Conclusion

We performed a fit of the short-distance couplings $\mathcal{C}_{7,9,10}$ appearing in the effective theory of ΔB=1 decays describing $b \to s\gamma$ and $b \to s\,\ell^+\ell^-$ transitions. Working in the SM-operator basis and assuming $\mathcal{C}_{7,9,10} \in \mathbb{R}$, we searched for hints of new physics in the simplest model-independent extension of the SM. For the first time, we included all relevant theory uncertainties in the analysis by means of 28 nuisance parameters. A total of 59 measurements of exclusive rare decays $B \to K^* \gamma$, $B \to K^{(*)} \ell^+ \ell^-$ and $B_s \to \mu^+ \mu^-$ obtained by CLEO, BaBar, Belle, CDF, and LHCb served as experimental inputs.

The main results of our analysis are the marginalized posterior distributions displayed in Fig. 7.1 and summarized in Tables 7.1 – 7.2. Using only subsets of the measurements, we performed several fits to highlight the impact of the individual decays. It is seen that the statistically most relevant contributions for $\mathcal{C}_7$ are from $B \to K^* \gamma$, and that the strongest constraints on $\mathcal{C}_9$ and $\mathcal{C}_{10}$ are from $B \to K^* \ell^+ \ell^-$ at high q^2. In all fits, multiple maxima due to discrete symmetries appear. Focusing on the fit with *all* experimental input, we observe two dominant solutions, the SM-like solution and the flipped-sign solution, arising from the approximate invariance of all observables under the transformation $\mathcal{C}_i \to -\mathcal{C}_i$. Within the Monte Carlo uncertainty, the two solutions have an equal posterior probability mass of roughly 52 % (SM) over 48 % (flipped-sign). Other local maxima exist, but their posterior masses are negligible. In each of the three 2D marginal posterior distributions of Fig. 7.1, the SM values $\mathcal{C}^{\mathrm{SM}}_{7,9,10}$ are close to the local maximum and inside the smallest 68 % region. Performing a back-of-the-envelope calculation, we placed lower limits on the energy scale of NP contributions; cf. Table 7.3. The resulting scales $\gtrsim 10\,\mathrm{TeV}$ are beyond the reach of direct production at the LHC.

Judging by two discrepancy variables, both solutions describe the data very well. For comparison, we considered the SM itself, by which we mean $\mathcal{C}_i$ fixed at SM predictions, but nuisance parameters allowed to vary with the same prior $P(\boldsymbol{\nu})$ that is used in the global fit. Then the SM also provided a very good fit of the data. The pull values in Fig. 7.5 and Fig. 7.6 show the individual distances from the best-fit point to the most likely experimental value for all inputs; we observed only moderate deviations.

Finally, we computed the Bayes factor to conduct model comparison; the data clearly favor the plain SM over a model with arbitrary real $\mathcal{C}_{7,9,10}$ — a tribute to Occam's razor. Apparently, adding extra model complexity through variable $\mathcal{C}_i$ does not yield a significantly better description of the data, hence the Bayes factor lends more support to the simpler model. Thus, from a purely statistical point of view, even the simplest model-independent extension of the SM is "too much" to describe the data at the current level of experimental and theory uncertainty. We emphasize that the presence of the sign-flipped solution still allows large NP contributions to the Wilson coefficients. However, the degeneracy within the set of considered observables does not allow us to distinguish them easily. This degeneracy is mildly broken by contributions of 4-quark operators, typically included in the effective Wilson coefficients $\mathcal{C}_{7,9} \to \mathcal{C}^{\mathrm{eff}}_{7,9}$. While the data do not favor one solution over the other, we have a much higher *a-priori* degree of belief in the SM-like solution given that the SM has been well tested to describe particle decays in and beyond the sector of B decays. Indeed, assuming improved theory uncertainties and current experimental central values in $B \to K^* \gamma$, the fit suggests that additional information on $\mathcal{C}^{\mathrm{eff}}_7$ may enhance the SM-like solution over the flipped-sign solution. For example, more information about $\mathcal{C}_7$ is available in the inclusive decay $B \to X_s \gamma$, where X_s represents any final state meson with one s quark.

For cross validation, we repeated the full fit with a second set of wide priors that may reflect a pessimist's opinion on the accuracy of form factors and the size of unknown subleading corrections — these uncertainties were enlarged by a factor of three.

Our comparison of the result from both sets in Fig. 7.2 indicates that the fit is stable and dominated by the experimental inputs encoded in the likelihood. Unsurprisingly, the marginal distributions do not coincide exactly; hence improvements both on the experimental side and on the theory side are desirable.

We observed that a fit with current $B \to K\ell^+\ell^-$ constraints prefers smaller values of $\mathcal{C}_{9,10}$ than a fit with the $B \to K^*\ell^+\ell^-$ constraints. Including both sets of constraints, the fit accommodated this tension by shifting the $B \to K$ form factors towards smaller values, and the $B \to K^*$ form factors towards larger values; see Fig. 7.1 and also Fig. 7.8(c), where the SM values of $(\mathcal{C}_9, \mathcal{C}_{10})$ are at the edge of the 95 % credibility region. While the tension shifted some marginal modes, it is important to stress that the posterior variance of *all* form factors was reduced significantly in the fit; cf. Fig. 7.3 and Fig. 7.4.

We computed updated predictions within the SM of selected observables in the angular distribution of $B \to K^*(\to K\pi)\ell^+\ell^-$; cf. Section 7.6 and Appendix B. In contrast to previous analyses [BHD10; BHD11b; APS12], we varied all parameters at once by performing uncertainty propagation (see Section 2.2) with the joint prior density $P(\boldsymbol{\nu})$. In this way, we properly treated the observables being nonlinear functions of $\boldsymbol{\nu}$ and further, we did not not implicitly restrict the outcome to a Gaussian distribution. All our $1\,\sigma$ intervals contain the central values of the previous analyses, and have similar width. We did not see an observable for which our interval is significantly larger; in most cases, the intervals are (slightly) smaller.

Based on the fit output, we predict ranges for currently unmeasured observables that exhibit a reduced form factor dependence. Given that the Wilson coefficients are *not* fixed in this case, it was a surprise that the predictions based on the fit output yielded smaller ranges than SM predictions based on prior knowledge; see Fig. 7.9. Evidently, the extra variance due to Wilson coefficients is more than compensated for by the reduced form factor uncertainties. This fact demonstrates the power of our statistical approach. For those observables where the impact of form factors is very small, the predictions were essentially identical; for example, compare $\left\langle H_T^{(3)} \right\rangle$ at high q^2 in Table 7.6 and Table B.2.

In the future, we want to improve the fit by including the measurements of the inclusive decays $B \to X_s\gamma$ as well as $B \to X_s\,\ell^+\ell^-$. In particular, $\mathcal{B}\,(B \to X_s\gamma)$ is known with a good relative precision of ~ 7 % both experimentally and theoretically; it has the potential to significantly reduce the posterior uncertainty on $\mathcal{C}_7$ by ~ 25 %. Furthermore, it aids in distinguishing the SM-like from the sign-flipped solution. Current $B \to X_s\,\ell^+\ell^-$ results from BaBar and Belle have a much higher uncertainty. This is unlikely to change before the end of data taking at the upcoming super flavor factories. Besides the inclusion of additional observables, further enhancements could arise when using the z parametrization of $B \to K^*$ form factors [BFW10]. Then we could continue to use LCSR at low q^2, and stabilize the extrapolation to high q^2 by fitting to new lattice results that are expected to appear in their final form in the near future [Win11] .

Apart from form factors, subleading corrections are the second-largest source of theory uncertainty. In our approach, they arise in two ways. The first source is the application of form factor relations to reduce the number of independent form factors (Section 5.3). If the form factors became available from the lattice at high q^2, we could trade the uncertainty arising from form factor relations for the (hopefully lower) uncertainty of the lattice form factors. Second, known subleading corrections $\propto (\Lambda/m_b)^2$ come into play due to dimension 5 operators in the OPE at high q^2 [BBF11]. Unfortunately, the additional form factors of these higher dimensional operators are unknown, but could in principle be calculated on the lattice as well. At low q^2, form factors can be calculated with the help of QCD sum rules that include subleading corrections usually omitted when applying large energy form factor relations. Furthermore, subleading corrections to the amplitude contributing to isospin breaking are known. These and other corrections given by Beneke, Feldmann, and Seidel [BFS05] are implemented in our software package EOS [Dyk+12].

On the data side, we see the following opportunities. The final analyses with the full data sample of Belle and CDF of $B \to K^{(*)}\ell^+\ell^-$ will likely have only a small impact. In contrast, new results from LHCb that include the full 2012 data set will contain roughly three times as many $B \to K^*\ell^+\ell^-$ events as the 2011 results we used in this work; those results, scheduled to be released to the public in 2013, might be accompanied by first analyses of ATLAS and CMS. In addition, there is the first LHCb analysis of $\mathcal{B}(B \to K\ell^+\ell^-)$ based on 2011 data that came out in the final stage of preparing this work [Aai+12b]. So far, the $B_s \to \mu^+\mu^-$ limit had a negligible impact on the fit, but if ATLAS, CMS, and LHCb continue to combine their search results also with the 2012 data, this channel could become competitive with $B \to K^*\ell^+\ell^-$ in constraining $|\mathcal{C}_{10}|$. In the long term after 2015, results from Belle II at SuperKEKB will have a substantial impact in all decay channels considered in the fit.

8 Conclusion

The primary goal of this thesis was to find signals of new physics in rare B meson decays mediated by $b \to s$ transitions. To this end, we considered the $\Delta B = 1$ effective field theory with the SM set of operators and three Wilson coefficients $\mathcal{C}_i, i = 7, 9, 10$ assumed real but otherwise unrestricted. This choice represents the minimal extension of the SM, as it assumes new physics only contributes to the Wilson coefficients. The Wilson coefficients were extracted in several global fits to a set of up to 59 measurements of observables in $B \to K^* \gamma$, $B \to K^{(*)} \ell^+ \ell^-$, and $B_s \to \mu^+ \mu^-$ reactions. Including all measurements, the fit revealed two degenerate regions of high posterior probability. The first solution contains the standard-model prediction in its $1\,\sigma$ region, and the second is related to the first through a simultaneous sign flip $\mathcal{C}_i \to -\mathcal{C}_i$; cf. Section 7.2.1.

Our choice to phrase the global fit in the Bayesian approach to probability theory proved crucial to the success of this work. As opposed to any other fit of rare B decays [BHD10; BHD11b; Des+11; Bob+12; APS12; AS12], we modeled all main uncertainties explicitly, using 28 nuisance parameters. While increasing the computational effort, this allowed us to consistently address the relevant questions. Most prominently, we performed a model comparison of the standard model versus new physics to see which is preferred by the data in Section 7.4. The Bayes factor yielded a relative change in probability of about 1500 in favor of the simpler SM. In addition, we investigated the goodness of fit in Section 7.3, and observed that the two fit solutions as well as the SM itself (nuisance parameters fitted) provided an excellent description of the data. Both facts combined — the Bayes factor and the good fit — imply that, at the current level of the rather high theoretical and experimental uncertainty, there is no extension that could outperform the SM when taking into account the extra model complexity.

Therefore, we refrained from analyzing nonminimal extensions such as $\mathcal{C}_i \in \mathbb{C}$ or new scalar and tensor operators. The situation may change in the near future, if LHCb measures $B \to K^* \ell^+ \ell^-$ observables optimized to reduce form factor dependence. In that case, the sensitivity to small deviations from SM expectations in the real and imaginary parts of the $\mathcal{C}_i$ would be greatly enhanced [BHD11a]. However, there are several other means by which we can improve the fit. Using more existing measurements, in particular, the inclusive decay $B \to X_s \gamma$, we can reduce the uncertainty on $\mathcal{C}_7$, and perhaps resolve the ambiguity between the two fit solutions. Also, ongoing efforts to compute $B \to K^{(*)} \ell^+ \ell^-$ form factors with lattice QCD aid in reducing the theory uncertainty. A more thorough discussion of the fit results and possible future enhancements is given in the conclusion of Chapter 7. The physics results described in this thesis are published [Bea+12].

The multimodality and the large dimensionality of the posterior posed a difficult numerical problem. Our solution was to combine Markov chain Monte Carlo and population Monte Carlo in order to draw samples from the posterior, benefiting from the speed-up of massive parallelization. Using the new combination successfully in all fits, we found the algorithm's performance very encouraging. However, there still is room for significant improvement; see Section 4.5 for a discussion of ideas in this direction.

In particular, outliers started to affect marginal distributions; the larger the parameter space, the stronger the effect. It remains to be seen whether the sampling can be enhanced to accomodate even more parameters in a future global fit with additional Wilson coefficients or extra nuisance parameters, for example to describe the uncertainties in $B \to X_s \gamma$.

Given the posterior samples, we calculated predictions based on the global fit output. In addition, we also presented SM predictions where we did not use any $b \to s$ measurements, but fixed the $\mathcal{C}_i$ to the SM values and varied the nuisance parameters according to the priors. Both sets of predictions were consistently obtained as a straightforward application of uncertainty propagation — another benefit of choosing the Bayesian approach. Within the SM, the smallest $1\,\sigma$ intervals are in good agreement with previous results. The big advantage of including the nuisance parameters in the fit became apparent in the prediction of standard observables such as $\mathcal{B}$ or F_L that exhibit form-factor dependence. As the fit constrained the form factors, the theory uncertainty on predictions was significantly reduced compared to the SM results; cf. Section 7.6.

In summary, we found no compelling evidence of new physics, yet at the same time the data allow large new-physics contributions due to the approximate invariance of observables under a simultaneous sign-flip of all Wilson coefficients. In the next few years, new results from LHCb, Belle II, and lattice QCD will hopefully enable us to disentangle the ambiguity; perhaps we will even be able to identify new physics in small deviations from the SM currently obscured by the fairly large uncertainties.

A Numerical input

The numerical values of those input parameters that enter the calculation of observables, but are not fitted to the data, are listed in Table A.1. Their impact on the fit uncertainty is small, either because they are known very precisely, or because they enter in numerically subleading contributions to the observables of interest.

The theory predictions of all the relevant semileptonic and radiative processes at large recoil are based on the QCDF results of [BFS01; BFS05]. These make use of the LCDA of the involved kaons which are parametrized in terms of Gegenbauer moments $a_n(M)$ ($M = K, K^*_\perp, K^*_\parallel$). In this work, we include terms in the expansion in Gegenbauer moments up to $n = 2$, using the central values in Table A.1.

Since the moments $a_n(M)$ also enter the computation of the $B \to K^*$ form factors via LCSR [BZ05a], variation of the former would lead to double counting. Furthermore, the residual influence of $a_n(M)$ on the observables is small compared to that of the other parameters considered. We therefore do not vary the Gegenbauer moments.

In addition, QCDF makes use of the decay constants f_M ($M = K, K^*_\perp, K^*_\parallel$) that enter in numerically suppressed contributions. The central values are listed in Table A.1.

The experimental inputs to the likelihood are listed for $B \to K^*\gamma$ (Table A.2), $B \to K\ell^+\ell^-$ (Table A.3), and $B \to K^*\ell^+\ell^-$ (Table A.4); for details on the definitions of observables, we refer to Chapter 6.

A.1 Nuisance parameters

In this section we present the nuisance parameters $\boldsymbol{\nu}$ that are included in the global fit to describe the main theory uncertainties. Priors are assumed independent a-priori, $P(\boldsymbol{\nu}) = \prod_i P(\nu_i)$, and clipped to finite ranges that correspond to the $3\,\sigma$ interval of the unclipped prior with the exception of the phases that naturally have a finite support (cf. Appendix A.1.3), where we do not perform any clipping. For the sake of readability, we categorize the individual nuisance parameters according to their impact.

A.1.1 Common nuisance parameters

The common nuisance parameters are those that enter most of the observables and are not specific to rare $b \to s$ decays. These are the parameters of the quark-mixing matrix $\boldsymbol{V}_{\text{CKM}}$ (5.9) and the b and c quark masses. Note that for computational reasons we omit the uncertainty related to the renormalization scale μ.

For the purpose of the fit, we take the CKM parameters from other observables such as tree decays. We parametrize the CKM matrix using the Wolfenstein parametrization (5.11) to $\mathcal{O}(\lambda^9)$ [Cha+05] and use the results of the tree-level fit of the UTfit collaboration [Bon+06] as priors in the fit of $b \to s$ decays. In this way, we include non-SM effects, but assume they do not affect tree-level decays. However, we use the results of the SM CKM fit in order to determine the uncertainties of observables in the framework of the SM in Section 7.6. Note that the CKM matrix elements only enter in the

$\alpha_s(m_Z)$	0.11762	[Nak+10]	m_μ	0.106 GeV	[Nak+10]
$\alpha_e(m_b)$	1/133	[Nak+10]	m_t^{pole}	173.3 GeV	[Tev09]
$\sin^2\theta_W$	0.23116	[Nak+10]	m_W	80.399 GeV	[Nak+10]
τ_{B^+}	1.638 ps	[Nak+10]	τ_{B^0}	1.525 ps	[Nak+10]
m_{B^+}	5.2792 GeV	[Nak+10]	m_{B^0}	5.2795 GeV	[Nak+10]
m_{K^+}	0.4937 GeV	[Nak+10]	m_{K^0}	0.4976 GeV	[Nak+10]
$m_{K^{*+}}$	0.8917 GeV	[Nak+10]	$m_{K^{*0}}$	0.8960 GeV	[Nak+10]
τ_{B_s}	1.472 ps	[Nak+10]	m_{B_s}	5.3663 GeV	[Nak+10]
$\lambda_{B,+}$	0.485 GeV	[BHD10]	$f_{B^{0,+}}$	0.212 GeV	[Sim+10]
f_K	0.1561 GeV	[Nak+10]			
$f_{K_\perp^*}(2\,\text{GeV})$	0.173 GeV	[BHD10]	$f_{K_\parallel^*}$	0.217 GeV	[BHD10]
$a_1(K)$	0.048	[BBL06]	$a_2(K)$	0.174	[BBL06]
$a_1(K_\perp^*)$	0.1	[BBL06]	$a_2(K_\perp^*)$	0.1	[BBL06]
$a_1(K_\parallel^*)$	0.1	[BBL06]	$a_2(K_\parallel^*)$	0.1	[BBL06]

Table A.1: The numerical input used in the analysis. The mass of the strange quark has been neglected throughout. τ_{B^0} (τ_{B^+}) denotes the lifetime of the neutral (charged) B meson. The following parameters appear in $\mathcal{A}(B \to (K, K^*)\,\ell^+\ell^-)$ at large recoil: $\lambda_{B,+}$ denotes the first inverse moment of the B-meson distribution amplitude, f_X is the decay constant of state X, and $a_{1,2}(M)$ are the first two Gegenbauer moments of the LCDA of the respective kaon states $M = K, K_\perp^*, K_\parallel^*$.

observable	value	ρ	
$\mathcal{B} \times 10^5$	$4.55^{+0.72}_{-0.68} \pm 0.34$		[Coa+00]
	$4.47 \pm 0.10 \pm 0.16$		[Aub+09]
	$4.01 \pm 0.21 \pm 0.17$		[Nak+04]
S	$-0.03 \pm 0.29 \pm 0.03$	5 %	[Aub+08]
C	$-0.14 \pm 0.16 \pm 0.03$		
S	$-0.32\,^{+0.36}_{-0.33} \pm 0.05$	8 %	[Ush+06]
C	$+0.20 \pm 0.24 \pm 0.05$		

Table A.2: Experimental results for CP-averaged $B^0 \to K^{*0}\gamma$ observables: branching fraction $\mathcal{B}$ (CLEO, BaBar, Belle) and time-dependent CP asymmetries S and C (BaBar, Belle), including the correlation coefficient ρ. Throughout, statistical errors are given first, followed by the systematic errors.

q^2-bin [GeV^2]	[1.00, 6.00]	[14.18, 16.00]	[> 16.00]	
$\langle\mathcal{B}\rangle \times 10^7$	$2.05\,^{+0.53}_{-0.48} \pm 0.07$	$1.46\,^{+0.41}_{-0.36} \pm 0.06$	$1.02\,^{+0.47}_{-0.42} \pm 0.06$	[Lee+12]
	$1.36\,^{+0.23}_{-0.21} \pm 0.08$	$0.38\,^{+0.19}_{-0.12} \pm 0.02$	$0.98\,^{+0.20}_{-0.18} \pm 0.06$	[Wei+09]
	$1.41 \pm 0.20 \pm 0.09$	$0.53 \pm 0.10 \pm 0.03$	$0.48 \pm 0.11 \pm 0.03$	[Aal+11a]

Table A.3: Experimental results for the CP-averaged branching fraction of charged $B^\pm \to K^\pm\mu^+\mu^-$ decays from BaBar [Lee+12], Belle [Wei+09], and CDF [Aal+11a], integrated over bins of q^2. The publicly available results of BaBar and Belle are unknown admixtures of charged and neutral B decays. The difference between interpreting the data as coming from either purely charged or purely neutral B decays is negligible [Bob+12]. The kinematic endpoint is $q^2_{max} = 19.21\,\mathrm{GeV}^2$.

q^2-bin [GeV^2]	[1.00, 6.00]	[14.18, 16.00]	[> 16.00]	
$\langle\mathcal{B}\rangle \times 10^7$	$2.05\,^{+0.53}_{-0.48} \pm 0.07$	$1.46\,^{+0.41}_{-0.36} \pm 0.06$	$1.02\,^{+0.47}_{-0.42} \pm 0.06$	[Lee+12]
	$1.49\,^{+0.45}_{-0.40} \pm 0.12$	$1.05\,^{+0.29}_{-0.26} \pm 0.08$	$2.04\,^{+0.27}_{-0.24} \pm 0.16$	[Wei+09]
	$1.42 \pm 0.41 \pm 0.08$	$1.34 \pm 0.26 \pm 0.08$	$0.97 \pm 0.26 \pm 0.06$	[Aal+11a]
	$2.10 \pm 0.20 \pm 0.20$	$1.08 \pm 0.13 \pm 0.07$	$1.32 \pm 0.15 \pm 0.09$	[Par12]
$\langle A_{\mathrm{FB}}\rangle$	$-0.02\,^{+0.18}_{-0.16} \pm 0.07$	$-0.31\,^{+0.19}_{-0.11} \pm 0.13$	$-0.34\,^{+0.26}_{-0.17} \pm 0.08$	[Poi12]
	$-0.26\,^{+0.30}_{-0.27} \pm 0.07$	$-0.70\,^{+0.22}_{-0.16} \pm 0.10$	$-0.66\,^{+0.16}_{-0.11} \pm 0.04$	[Wei+09]
	$-0.36\,^{+0.28}_{-0.46} \pm 0.11$	$-0.40\,^{+0.21}_{-0.18} \pm 0.07$	$-0.66\,^{+0.26}_{-0.18} \pm 0.19$	[Aal+12]
	$0.18 \pm 0.06\,^{+0.02}_{-0.01}$	$-0.49\,^{+0.06}_{-0.04}\,^{+0.05}_{-0.02}$	$-0.30 \pm 0.07\,^{+0.01}_{-0.04}$	[Par12]
$\langle F_L\rangle$	$0.47 \pm 0.13 \pm 0.04$	$0.42\,^{+0.12}_{-0.16} \pm 0.11$	$0.47\,^{+0.18}_{-0.20} \pm 0.13$	[Poi12]
	$0.67 \pm 0.23 \pm 0.05$	$-0.15\,^{+0.27}_{-0.23} \pm 0.07$	$0.12\,^{+0.15}_{-0.13} \pm 0.02$	[Wei+09]
	$0.60\,^{+0.21}_{-0.23} \pm 0.09$	$0.32 \pm 0.14 \pm 0.03$	$0.16\,^{+0.22}_{-0.18} \pm 0.06$	[Aal+12]
	$0.66 \pm 0.06\,^{+0.04}_{-0.03}$	$0.35\,^{+0.07}_{-0.06}\,^{+0.07}_{-0.02}$	$0.37\,^{+0.06}_{-0.07}\,^{+0.03}_{-0.04}$	[Par12]
$\left\langle A_T^{(2)}\right\rangle$	$1.6\,^{+1.8}_{-1.9} \pm 2.2$	$0.4 \pm 0.8 \pm 0.2$	$-0.9 \pm 0.8 \pm 0.4$	[Aal+12]
$\langle 2S_3\rangle$	$0.10\,^{+0.15}_{-0.16}\,^{+0.02}_{-0.01}$	$0.04\,^{+0.15}_{-0.19}\,^{+0.04}_{-0.02}$	$-0.47\,^{+0.21}_{-0.10}\,^{+0.03}_{-0.05}$	[Par12]

Table A.4: Experimental results of $B^0 \to K^{*0}\ell^+\ell^-$ for the CP-averaged branching fraction $\mathcal{B}$, lepton forward-backward asymmetry A_{FB}, longitudinal K^*-polarization fraction F_L, the transversity observable $A_T^{(2)}$ and $(2S_3)$ from BaBar [Lee+12; Poi12], Belle [Wei+09], CDF [Aal+11a; Aal+12], and LHCb [Par12]. Note that the sign of A_{FB} is reversed due to a different definition of θ_ℓ in the experimental community. The kinematic endpoint is $q^2_{max} = 22.86\,\mathrm{GeV}^2$.

A	0.804 ± 0.010	[Bon+06]	λ	0.22535 ± 0.00065	[Bon+06]
$\bar{\rho}$	0.111 ± 0.070	[Bon+06]	$\bar{\eta}$	0.381 ± 0.030	[Bon+06]
$m_c(\mu = m_c)$	$(1.27\,^{+0.07}_{-0.09})$ GeV	[Nak+10]	$m_b(\mu = m_b)$	$(4.19\,^{+0.18}_{-0.06})$ GeV	[Nak+10]

Table A.5: Common nuisance parameters. The CKM Wolfenstein parameter values as obtained from the CKM *tree-level fit*; cf. Sec. Section 2.3 and (5.11). Quark masses are given in the $\overline{\text{MS}}$ scheme.

	r_1	r_2	m_R^2 [GeV^2]	m_{fit}^2 [GeV^2]
V	0.923	−0.511	5.32^2	49.40
A_1	–	0.290	–	40.38
A_2	−0.084	0.343	–	52.00

Table A.6: The parameters of the form factors V and$A_{1,2}$ are defined in (A.1) and (A.2).

combinations $V_{tb}V_{ts}^*$ and $V_{ub}V_{us}^*$. Although numerically negligible, the latter included in the analysis. It becomes relevant only for CP-asymmetric observables. All priors are Gaussian, with their $1\,\sigma$ ranges given in Table A.5.

The values of the quark masses m_b and m_c enter most observables. In order to account for the asymmetric errors, we use LogGamma distributions (see Section A.2) as priors whose modes and 68 %-probability intervals match the values given in Table A.5.

A.1.2 $B \to K^{(*)}$ form factors and f_{B_s}

The heavy-to-light form factors $f_{+,T,0}$ for $B \to K$ as well as V, $A_{0,1,2}$, and $T_{1,2,3}$ for $B \to K^*$ transitions present a major source of uncertainty in predictions of rare exclusive B decays. They are functions of the dilepton invariant mass q^2 and we adopt the definition used in [BF01; BFS05; Kho+10; BZ05a]. Due to the application of form factor relations at large and low recoil, only f_+ enters $B \to K$, and only V and $A_{1,2}$ enter $B \to K^*$ transitions[1]. The application of form factor relations introduces uncertainties of order Λ_{QCD}/m_b that will be discussed in Appendix A.1.3.

Currently, the form factors are only known from LCSR which are applicable at low q^2. Lattice QCD can provide results at high q^2, where quenched results of some form factors [BLM07; AlH+10] are available and some preliminary unquenched results have been reported [Liu+09; Zho+11; Liu+11]. An extensive discussion of the q^2-shape parametrization using series expansion and a fit to low-q^2 LCSR combined with high-q^2 lattice results (where available) is given in [BFW10].

With regard to $B \to K^*$ form factors $V, A_{1,2}$, we use the LCSR results at low q^2 as given in [BZ05a], where the extrapolation to high-q^2 is based on a (multi-)pole ansatz

$$V = \frac{r_1}{1 - q^2/m_R^2} + \frac{r_2}{1 - q^2/m_{\text{fit}}^2}, \tag{A.1}$$

$$A_1 = \frac{r_2}{1 - q^2/m_{\text{fit}}^2}, \qquad A_2 = \frac{r_1}{1 - q^2/m_{\text{fit}}^2} + \frac{r_2}{(1 - q^2/m_{\text{fit}}^2)^2}, \tag{A.2}$$

[1] The form factors f_0 and A_0 do not contribute within the framework of the SM operator basis, up to negligible terms suppressed by m_ℓ^2/q^2.

and the numerical values of the parameters given in Table A.6. We do not vary these parameters themselves as they strongly depend on the LCSR analysis, but rather assign one multiplicative scaling factor ζ_i per form factor ($i = V, A_1, A_2$) to model the respective uncertainty such that the value $\zeta_i = 1.0$ corresponds to the central value of the form factor. What do we know about ζ_i? The only information that we use is that the current consensus in the community is that the form factors are known to an accuracy of (10 – 15) %. Based on the MAXENT principle and (2.26), we assign a Gaussian prior with a width of $\sigma = 0.15$ (i.e., 15 % uncertainty) and support extending up to $3\,\sigma$ (i.e., at most 45 % uncertainty) (see Table A.7); e.g.,

$$V \to \zeta_V V, \qquad P(\zeta_V) = \mathcal{N}(\zeta_V|\mu = 1, \sigma = 0.15)\,. \tag{A.3}$$

Note that in this way we do not vary the q^2 shape of the form factors. At large recoil, two universal form factors [BFS05] appear:

$$\xi_\perp \equiv \frac{m_B}{m_B + m_{K^*}} V\,, \qquad \xi_\parallel \equiv \frac{m_B + m_{K^*}}{2E_{K^*}} A_1 - \frac{m_B - m_{K^*}}{m_B} A_2\,. \tag{A.4}$$

Their variation is obtained by the uncorrelated variation of V and $A_{1,2}$ as described above.

Since we calculate the $B \to K^*\gamma$ matrix element within QCDF for $q^2 = 0$, all nuisance parameters that affect the process $B \to K^*\,\ell^+\ell^-$ in the large recoil region likewise affect the radiative process, as far as they are applicable.

Note that the prediction of the form-factor independent ratio T_1/V differs between heavy quark symmetry and LCSR. Favoring heavy quark symmetry, we introduce a correction factor 0.319464 to reduce T_1 in order to make both approaches consistent. The number differs slightly from the original proposal in [BFS01, Eq. (46)] as we use more recent input numbers.

With regard to the $B \to K$ form factor f_+, we use the BCL parametrization [BCL09] of the LCSR results [Kho+10]

$$f_+(q^2) = \frac{f_+(0)}{1 - q^2/m^2_{\text{res},+}} \left[1 + b_1^+ \left(z(q^2) - z(0) + \frac{1}{2}\left[z(q^2)^2 - z(0)^2\right]\right)\right], \tag{A.5}$$

$$z(s) = \frac{\sqrt{\tau_+ - s} - \sqrt{\tau_+ - \tau_0}}{\sqrt{\tau_+ - s} - \sqrt{\tau_+ - \tau_0}}, \qquad \tau_0 = \sqrt{\tau_+}\left(\sqrt{\tau_+} - \sqrt{\tau_+ - \tau_-}\right), \qquad \tau_\pm = (m_B \pm m_K)^2\,.$$

This parametrization depends on the central value of the form factor at $q^2 = 0$, $f_+(0)$, and the slope parameter b_1^+ (and $m_{\text{res},+} = 5.412\,\text{GeV}$). At large recoil, the dipole form factor f_T is replaced by the large-energy universal form factor $\xi_P \equiv f_+$ [BF01; BHP07]. At low recoil, the dipole form factor f_T is substituted for by means of the improved Isgur-Wise relation [BHD11b]. We assign LogGamma priors to $f_+(0)$ and b_1^+ with uncertainties listed in Table A.7.

In addition, we vary the decay constant f_{B_s} of the B_s meson, since it constitutes the dominant uncertainty in the decay $B_s \to \mu^+\mu^-$. The most recent lattice results [McN+12; Baz+12] have been averaged [LLVdW10], yielding the number listed in Table A.7.

The prior elicitation requires a bit of educated guessing; in order to assess the dependence of the fit on the choice of priors quantitatively, we adopt two sets of priors and repeat the fit. The first set reflects the uncertainties as reported by the authors of [BZ05a; Kho+10; LLVdW10], thereby assuming the extrapolation of form factors to high q^2 has the same uncertainties as predicted by LCSR at low q^2. In the second set, we triple the uncertainties. Both sets are given in Table A.7.

parameter	central	nominal		wide	
		$1\,\sigma$	support	$1\,\sigma$	support
ζ_{V,A_1,A_2}	1.0	0.15	$3\,\sigma$	0.45	$3\,\sigma$
$f_+(0)$	0.34	$[0.32, 0.39]$	$[0.28, 0.49]$	$[0.28, 0.49]$	$[0.0, 0.79]$
b_1^+	-2.1	$[-3.7, -1.2]$	$[-6.9, 0.6]$	$[-6.9, 0.6]$	$[-10, 3.7]$
f_{B_s}	227.7 MeV	6.2 MeV	$3\,\sigma$	18.6 MeV	$3\,\sigma$
$\zeta_{K^*}^{ij}, \zeta_K$	1.0	0.15	$3\,\sigma$	0.45	$[0.0, 2.0]$
$\lvert r_{0,\perp,\parallel}\rvert, \lvert r_K\rvert$	0.0	0.15	$3\,\sigma$	0.45	$3\,\sigma$

Table A.7: Priors of the nuisance parameters of the $B \to K^{(*)}$ form factors, the B_s decay constant f_{B_s}, and parametrization of lacking subleading corrections at low q^2 ($i = L, R$ and $j = 0, \perp, \parallel$) and high q^2, specified for the nominal and wide set. All priors are Gaussian except those for $f_+(0)$ and b_1^+; we give the central value, the $1\,\sigma$ ranges, and the support of the prior. The nominal $1\,\sigma$ ranges of V and $A_{1,2}$ correspond to uncertainties quoted in [BZ05a], whereas, $f_+(0)$ and b_1^+ are taken from the LCSR analysis [Kho+10]; however, possible correlations among $f_+(0)$ and b_1^+ are not available. The uncertainty of f_{B_s} is due to the combined uncertainties of [McN+12] and [Baz+12].

A.1.3 Subleading Λ/m_b corrections

There are several distinct sources of Λ/m_b corrections arising in exclusive $B \to K^{(*)}\ell^+\ell^-$ decays. Here Λ is assumed to be of the order of the strong scale, however the particular physical meaning depends on the framework. When using power counting, we use the generic value of 500 MeV.

The first type is due to the form factor relations in the limit of heavy quark masses [IW90], which is valid in the entire kinematic region of q^2. At the leading order in Λ/m_b, they relate the $B \to K^*$ ($B \to K$) tensor form factors $T_{1,2,3}$ (f_T) to vector and axial-vector V and $A_{1,2}$ (f_+) form factors[2]. This approximation receives a further numerical suppression due to $\mathcal{C}_7/\mathcal{C}_9 \sim \mathcal{O}(0.1)$. The additional collinear limit [Cha+99; BF01] at low q^2 allows us to eliminate another $B \to K^*$ form factor, introducing an additional subleading uncertainty not suppressed by $\mathcal{C}_7/\mathcal{C}_9$. Besides subleading corrections due to the use of form factor relations, the two distinct expansions in Λ/m_b, QCDF at low q^2 and the OPE at high q^2, introduce a second type at the amplitude level, when truncating the expansion after the leading order in Λ/m_b.

At low q^2, QCDF (or equivalently soft collinear effective theory) provides a possibility to calculate such corrections, which are in general suppressed by a factor of Λ/m_b. But in some subleading corrections one encounters infrared divergences [FM03]. In principle, the partially known corrections [FM03; BFS05] could be included as an estimate of the lacking corrections, but here we model them by 6 real scale factors for each of the transversity amplitudes $A_{\perp,\parallel,0}^{L,R}$ in the case of $B \to K^*\ell^+\ell^-$ and one for $B \to K\,\ell^+\ell^-$. These scale factors $\zeta_{K^*}^{ij}$ ($i = L, R$ and $j = 0, \perp, \parallel$) and ζ_K are included in the fit with a Gaussian prior $P(\cdot) = \mathcal{N}(\cdot|1, 0.15)$ as in (A.3). For the nominal priors, the support extends up to $3\,\sigma$. A $1\,\sigma$ range of $0.45 \approx \Lambda/m_b$ with a support $[0.0, 2.0]$ is chosen for the

[2] The authors of [Alt+09] suggest that such corrections can be accounted for at low q^2, if form factor relations are not used in the leading-order contribution (in Λ/m_b and α_s) to the amplitude.

wide-prior scenario.

At high q^2, the interaction of the 4-quark operators and the electromagnetic current, which couples to the pair of leptons, is treated within a local operator product expansion either in full QCD [BBF11] or with subsequent matching on heavy quark effective theory (HQET) [GP04]. In both approaches, subleading corrections to the decay amplitudes arise at $(\Lambda/m_b)^2$ and $\alpha_s\,\Lambda/m_b$, respectively, which are of similar numerical size. The additional suppression factor of Λ/m_b or α_s, yields smaller theory uncertainties due to omission of subleading corrections at high q^2 in contrast to the low-q^2 region. This is also not spoiled by the use of form factor relations [GP04; BHD10] for tensor form factors $T_{1,2,3}$ (f_T) due to the accompanying numerical suppression by $\mathcal{C}_7/\mathcal{C}_9$, which depends on the new physics contributions. Note that for both approaches, full QCD and HQET, the subleading corrections are known in part, and in the future it is conceivable that they can be included completely. For example, the unknown subleading form factor arising in [BBF11] could be calculated on the lattice. We follow [GP04], using $\alpha_s(m_b) \sim 0.3$. This gives rise to 3 complex contributions $r_a \sim \Lambda/m_b$ $(a = 0, \perp, \|)$ for $B \to K^*\ell^+\ell^-$ [BHD11b] and one complex contribution $r_K \sim \Lambda/m_b$ for $B \to K\,\ell^+\ell^-$ [Bob+12], which are additive at the amplitude level. We treat the complex-valued subleading contributions r_a with eight additional real-valued nuisance parameters, assigning Gaussian priors with central value 0, a $1\,\sigma$ range of $0.15 \approx \Lambda/m_b$, and a support up to $3\,\sigma$ to describe $|r_i|$. Invoking MAXENT and (2.27), we describe our state of knowledge of the accompanying phases $\arg r_i$ with uniform priors on $[-\pi/2, \pi/2]$. A tripled $1\,\sigma$ range of $0.45 \approx \Lambda/m_b$ and a support up to $3\,\sigma$ is chosen for the wide-prior scenario.

The prior choices are summarized in Table A.7.

A.2 LogGamma distribution

Consider a parameter x whose reported uncertainties are asymmetric, $x = \mu^{+\sigma_+}_{-\sigma_-}, \sigma_- \neq \sigma_+$. In this case, we use the LogGamma distribution [Cro10] to obtain a continuous prior over the given range of x. The LogGamma family is a continuous unimodal three-parameter family of probability distributions

$$\begin{aligned} \mathrm{LogGamma}(x|l,\lambda,\alpha) &= \frac{1}{\Gamma(\alpha)|\lambda|}\exp\left(\alpha\left(\frac{x-l}{\lambda}\right) - \exp\left(\frac{x-l}{\lambda}\right)\right) \\ &\text{for } x,\ l,\ \lambda,\ \alpha, \text{ in } \mathbb{R},\ \alpha > 0, \\ &\text{support } -\infty \le x \le \infty. \end{aligned} \tag{A.6}$$

The three parameters are uniquely fixed by demanding that the mode of $P(x)$ be at μ, that the interval $[\mu - \sigma_-, \mu + \sigma_+]$ contain 68 %, and that the density be identical at $\mu - \sigma_-$ and $\mu + \sigma_+$. More concisely, the three conditions are:

$$\arg\max_x\ P(x) = \mu \tag{A.7}$$

$$\int_{\mu-\sigma_-}^{\mu+\sigma_+} \mathrm{d}x\, P(x) = 0.68 \tag{A.8}$$

$$P(\mu - \sigma_-) = P(\mu + \sigma_+). \tag{A.9}$$

For a finite range of x, say $[x_{min}, x_{max}]$, the resulting density is normalized such that $\int_{x_{min}}^{x_{max}} \mathrm{d}x\, P(x) = 1$. While (A.7) is used to fix the location parameter $l = \mu - \lambda \log \alpha$,

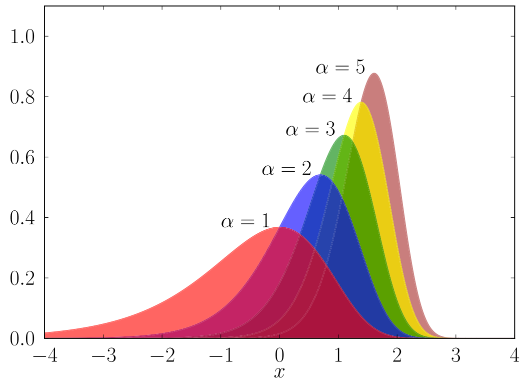


Figure A.1: The $\text{LogGamma}(x|\mu = 0, \lambda = 1, \alpha)$ density for various values of the shape parameter α.

the scale parameter λ and the shape parameter α must be extracted numerically by solving the coupled equations (A.8) and (A.9). Note that the asymmetry is governed by α: $\text{LogGamma}(\cdot)$ approaches a symmetric Gaussian distribution in the limit $\alpha \to \infty$, whereas the skew diverges and the mode is shifted to ∞ as $\alpha \to 0$; see Fig. A.1 for a graphical illustration.

Solving the constraints (A.8) and (A.9) numerically is not an easy task, because the PDF (A.6) depends strongly on the scale and shape parameter. It is therefore imperative to start a gradient-based numerical optimization at a very good starting point. In the following, we explain how to determine that starting point. Without loss of generality, let $\sigma_+ > \sigma_-$ and define the *normalized* uncertainties as $\hat{\sigma}_+ \equiv \sigma_+/\sigma_-$, $\hat{\sigma}_- \equiv 1$. The normalized scale factor, $\hat{\lambda}$, is then negative and related to the actual scale λ as $\lambda = \sigma_-\hat{\lambda}$. It turns out that the initial values of $\hat{\lambda}$ and α can be chosen independently as functions of $\hat{\sigma}_+$. Our empirical results are

$$\hat{\lambda}_0 = -56 + 55\Phi(\hat{\sigma}_+ - 1|\mu = 0, \sigma = 0.05) \tag{A.10}$$

$$\alpha_0 = \left(\frac{1.13}{\hat{\sigma}_+ - 1}\right)^{1.3}, \tag{A.11}$$

with the Gaussian cumulative function $\Phi(\cdot|\mu, \sigma)$. In our analysis, these values differed from the output of the optimization within $5\,\%(\alpha)$, and $10\,\%(\lambda)$, respectively. Note that this procedure does *not* yield meaningful results for very symmetric uncertainties where $\hat{\sigma}_+ < 1.06$. In that case, we recommend using a Gaussian instead to avoid the singularity as $\alpha \to \infty$. Similarly, the procedure becomes unstable for $\hat{\sigma}_+ \gg 1$.

B Standard model predictions

In this appendix we provide q^2-integrated SM predictions for measured and unmeasured observables, focusing on those low- and high-q^2 bins that are currently used in experimental analyses and are also accessible to theoretical methods. All quantities are CP averaged and lepton-mass effects have been taken into account using $\ell = \mu$. The theory uncertainties are calculated using uncertainty propagation (Section 2.2) with the (nominal) prior distributions of the 28 nuisance parameters presented in Appendix A.1. Note that the uncertainty on the renormalization scale μ is not incorporated.

The results are listed in Table B.1 and Table B.2 for low and high q^2 in the form

$$X = X^{*}\,{}^{+\sigma_+}_{-\sigma_-}\,(X_{\text{central}})\ , \tag{B.1}$$

where X^* is the mode of resulting distribution $P(X)$ approximated by a histogram, $[X^* - \sigma_-, X^* + \sigma_+]$ is the minimal interval containing at least 68 %, and X_{central} is the value obtained by evaluating X at the prior mode, corresponding to the conventional estimate of the SM prediction. Some example distributions are displayed in Fig. 7.9.

At low q^2, we do not predict J_3, J_9 and associated optimized observables $A_T^{(2)}$ and $A_T^{(\text{im})}$, since they vanish at leading order in QCDF (including the α_s corrections), although we obtain non-vanishing values due to the implementation of subleading terms of kinematic origin ($\sim M_{K^*}/M_B$).

At high q^2, J_7, J_8, and J_9 are zero at leading order in the OPE and when applying form factor relations, so is $A_T^{(\text{im})}$. Furthermore, we recall that F_L and $A_T^{(2,3)}$ become short-distance independent [BHD10] within the framework of the SM operator basis, and predictions are strongly dependent on the extrapolation of the form factor results from low q^2 obtained using LCSR.

We do not predict J_{6c} since it vanishes in the absence of scalar and tensor operators.

Observable	[2.0, 4.3]	[1.0, 6.0]
$\langle \mathcal{B}_K \rangle \times 10^7$ †	$0.85\,^{+0.25}_{-0.13}\,(0.81)$	$1.85\,^{+0.54}_{-0.28}\,(1.75)$
$\langle \mathcal{B}_{K^*} \rangle \times 10^7$ ‡	$0.69\,^{+0.77}_{-0.41}\,(1.05)$	$1.64\,^{+1.80}_{-0.83}\,(2.46)$
$\langle A_{\mathrm{FB}} \rangle$	$0.055\,^{+0.087}_{-0.033}\,(0.086)$	$0.03\,^{+0.07}_{-0.02}\,(0.05)$
$\langle F_L \rangle$	$0.85\,^{+0.08}_{-0.20}\,(0.78)$	$0.81\,^{+0.09}_{-0.22}\,(0.73)$
$\langle J_{1s} \rangle \times 10^8$	$1.18\,^{+0.48}_{-0.35}\,(1.26)$	$3.43\,^{+1.37}_{-0.95}\,(3.66)$
$\langle J_{1c} \rangle \times 10^7$	$0.31\,^{+0.57}_{-0.29}\,(0.63)$	$0.83\,^{+1.07}_{-0.76}\,(1.37)$
$\langle J_{2s} \rangle \times 10^8$	$0.39\,^{+0.16}_{-0.12}\,(0.42)$	$1.13\,^{+0.45}_{-0.31}\,(1.21)$
$\langle J_{2c} \rangle \times 10^7$	$-0.30\,^{+0.28}_{-0.56}\,(-0.61)$	$-0.79\,^{+0.75}_{-1.05}\,(-1.33)$
$\langle J_4 \rangle \times 10^8$	$0.57\,^{+0.39}_{-0.24}\,(0.77)$	$1.43\,^{+0.82}_{-0.62}\,(1.82)$
$\langle J_5 \rangle \times 10^8$	$-0.69\,^{+0.37}_{-0.64}\,(-1.07)$	$-1.80\,^{+0.88}_{-1.37}\,(-2.58)$
$\langle J_{6s} \rangle \times 10^8$	$0.84\,^{+0.45}_{-0.29}\,(0.90)$	$1.19\,^{+0.87}_{-0.74}\,(1.21)$
$\langle J_7 \rangle \times 10^9$	$2.52\,^{+1.50}_{-1.06}\,(2.78)$	$5.86\,^{+3.03}_{-2.62}\,(6.21)$
$\langle J_8 \rangle \times 10^9$	$-0.89\,^{+0.49}_{-0.57}\,(-0.97)$	$-1.79\,^{+0.94}_{-1.36}\,(-2.14)$
$\langle A_T^{(3)} \rangle$	$0.45\,^{+0.12}_{-0.08}\,(0.50)$	$0.42\,^{+0.11}_{-0.08}\,(0.47)$
$\langle A_T^{(4)} \rangle$	$0.63\,^{+0.17}_{-0.17}\,(0.69)$	$0.64\,^{+0.18}_{-0.15}\,(0.71)$
$\langle A_T^{(5)} \rangle$	$0.41\,^{+0.03}_{-0.05}\,(0.42)$	$0.48\,^{+0.01}_{-0.03}\,(0.48)$
$\langle A_T^{(\mathrm{re})} \rangle$	$0.61\,^{+0.10}_{-0.13}\,(0.54)$	$0.29\,^{+0.14}_{-0.14}\,(0.25)$
$\langle H_T^{(1)} \rangle$	$0.45\,^{+0.08}_{-0.08}\,(0.48)$	$0.42\,^{+0.07}_{-0.07}\,(0.45)$
$\langle H_T^{(2)} \rangle$	$-0.29\,^{+0.08}_{-0.08}\,(-0.34)$	$-0.29\,^{+0.07}_{-0.07}\,(-0.33)$

Table B.1: SM predictions of q^2-integrated observables at *low* q^2 in the bins $q^2 \in [q^2_{\mathrm{min}}, q^2_{\mathrm{max}}]$ for † $B^- \to K^-\mu^+\mu^-$ and ‡ $\bar{B}^0 \to \bar{K}^{*0}\mu^+\mu^-$. We list the mode and the smallest 68 % interval of the probability distribution, along with the value obtained by the conventional method of setting all nuisance parameters to the prior modes (in parentheses).

Observable	[14.18, 16.0]	[> 16.0]	[> 14.18]
$\langle \mathcal{B}_K \rangle \times 10^7$ †	$0.39^{+0.22}_{-0.09}$ (0.37)	$0.73^{+0.43}_{-0.22}$ (0.68)	$1.11^{+0.66}_{-0.28}$ (1.04)
$\langle \mathcal{B}_{K^*} \rangle \times 10^7$ ‡	$1.19^{+0.37}_{-0.31}$ (1.26)	$1.41^{+0.40}_{-0.38}$ (1.46)	$2.57^{+0.80}_{-0.68}$ (2.72)
$\langle A_{\rm FB} \rangle$	$-0.44^{+0.07}_{-0.07}$ (−0.44)	$-0.37^{+0.06}_{-0.07}$ (−0.38)	$-0.40^{+0.06}_{-0.07}$ (−0.41)
$\langle F_L \rangle$	$0.38^{+0.04}_{-0.06}$ (0.36)	$0.35^{+0.02}_{-0.03}$ (0.34)	$0.36^{+0.04}_{-0.05}$ (0.35)
$\langle J_{1s} \rangle \times 10^8$	$4.44^{+0.96}_{-1.00}$ (4.51)	$5.10^{+1.48}_{-1.11}$ (5.44)	$9.70^{+2.31}_{-2.21}$ (9.96)
$\langle J_{1c} \rangle \times 10^8$	$3.23^{+1.31}_{-1.37}$ (3.43)	$3.40^{+1.41}_{-1.07}$ (3.72)	$6.64^{+2.75}_{-2.43}$ (7.14)
$\langle J_{2s} \rangle \times 10^8$	$1.48^{+0.32}_{-0.33}$ (1.50)	$1.70^{+0.49}_{-0.37}$ (1.81)	$3.23^{+0.77}_{-0.74}$ (3.31)
$\langle J_{2c} \rangle \times 10^8$	$-3.21^{+1.36}_{-1.31}$ (−3.41)	$-3.38^{+1.07}_{-1.41}$ (−3.70)	$-6.61^{+2.42}_{-2.74}$ (−7.11)
$\langle J_3 \rangle \times 10^8$	$-0.99^{+0.59}_{-0.71}$ (−1.11)	$-2.12^{+0.89}_{-0.82}$ (−2.19)	$-3.06^{+1.44}_{-1.57}$ (−3.29)
$\langle J_4 \rangle \times 10^8$	$2.47^{+0.95}_{-0.85}$ (2.65)	$3.10^{+1.08}_{-0.96}$ (3.27)	$5.49^{+2.06}_{-1.77}$ (5.92)
$\langle J_5 \rangle \times 10^8$	$-3.36^{+0.87}_{-0.87}$ (−3.54)	$-2.95^{+0.63}_{-0.80}$ (−3.17)	$-6.23^{+1.34}_{-1.79}$ (−6.72)
$\langle J_{6s} \rangle \times 10^7$	$-0.52^{+0.10}_{-0.12}$ (−0.55)	$-0.53^{+0.11}_{-0.12}$ (−0.56)	$-1.05^{+0.22}_{-0.24}$ (−1.11)
$\langle A_T^{(2)} \rangle$	$-0.38^{+0.17}_{-0.18}$ (−0.37)	$-0.64^{+0.15}_{-0.10}$ (−0.60)	$-0.51^{+0.16}_{-0.16}$ (−0.50)
$\langle A_T^{(3)} \rangle$	$1.45^{+0.29}_{-0.31}$ (1.47)	$1.95^{+0.42}_{-0.40}$ (2.01)	$1.67^{+0.36}_{-0.34}$ (1.72)
$\langle A_T^{(4)} \rangle$	$0.66^{+0.14}_{-0.14}$ (0.67)	$0.48^{+0.10}_{-0.10}$ (0.48)	$0.56^{+0.12}_{-0.11}$ (0.57)
$\langle A_T^{(5)} \rangle$	$0.085^{+0.008}_{-0.008}$ (0.081)	$0.111^{+0.014}_{-0.014}$ (0.109)	$0.123^{+0.012}_{-0.012}$ (0.120)
$\langle A_T^{(\rm re)} \rangle$	$-0.982^{+0.110}_{-0.003}$ (−0.915)	$-0.777^{+0.099}_{-0.089}$ (−0.767)	$-0.843^{+0.075}_{-0.087}$ (−0.834)
$\langle H_T^{(1)} \rangle$	$0.9996^{+0.0002}_{-0.0003}$ (0.9996)	$0.9986^{+0.0008}_{-0.0007}$ (0.9986)	$0.9970^{+0.0017}_{-0.0018}$ (0.9969)
$\langle H_T^{(2)} \rangle$	$-0.9844^{+0.0027}_{-0.0020}$ (−0.9853)	$-0.9719^{+0.0034}_{-0.0024}$ (−0.9722)	$-0.9748^{+0.0040}_{-0.0031}$ (−0.9751)
$\langle H_T^{(3)} \rangle$	$-0.9837^{+0.0024}_{-0.0018}$ (−0.9845)	$-0.9614^{+0.0017}_{-0.0011}$ (−0.9618)	$-0.9606^{+0.0018}_{-0.0016}$ (−0.9613)

Table B.2: SM predictions of q^2-integrated observables at *high* q^2 in the bins $q^2 \in [q^2_{\rm min}, q^2_{\rm max}]$ for †$B^- \to K^-\mu^+\mu^-$ and ‡$\bar{B}^0 \to \bar{K}^{*0}\mu^+\mu^-$. We list the mode and the smallest 68 % interval of the probability distribution, along with the value obtained by the conventional method of setting all nuisance parameters to the prior modes (in parentheses).

C Goodness of fit

In the ideal case, it is possible to calculate the degree of belief in a model based on the data. This option is only available when a complete set of models and their prior probabilities can be defined. However, the conditions necessary for this ideal case are usually not met in practice. We nevertheless often want to make some statement concerning the validity of the model(s). We then are left with using probabilities of data outcomes assuming the model to try to make some judgments. These probabilities can be determined deductively since the model is assumed, and therefore frequencies of possible outcomes can be produced within the context of the model. These can then be used to produce frequency distributions of discrepancy variables (defined below), and p values (defined below) can be calculated using the distributions and the observed values. The use of p values has been widely discussed in the literature (see, e.g., [BB00; SBB01]) and many authors have commented that p values are frequently misused in claiming support for models, cf. [Sch96; Blo+06]. We give a Bayesian argumentation for the use of p values to make judgments on model validity, and it is in this Bayesian sense that we will use p values.

C.1 General approach

For a given model, we can define one or more discrepancy variables — scalar functions of the data — and calculate its expected frequency distribution *assuming the model*. We use $R(\boldsymbol{x}|\boldsymbol{\theta}, M)$ and $R(D|\boldsymbol{\theta}, M)$ to denote discrepancy variables evaluated with a possible set of observations $\boldsymbol{x}$ for given model and parameter values, and for the observed data, $\boldsymbol{x} = D$, respectively. To simplify the notation, we will occasionally drop the arguments on R and use R^D to denote the value of the discrepancy variable found from the data set at hand. R can be interpreted as a random variable (e.g., possible χ^2 values for a given model), whereas R^D has a fixed value (e.g., the observed χ^2 derived from the data set at hand). If the discrepancy variable is well chosen, then the distribution for a "good" model should look significantly different than for a "bad" model. Finding the discrepancy variable in the region populated by incorrect models then gives us cause to think our model is not adequate.

C.2 p value

A p value is the probability that, in a future experiment, the discrepancy variable will have a larger value (indicating greater deviation of the data from the model) than the value observed, assuming that the model is correct and all experimental effects are perfectly known. In other words, not only is the model the correct one to describe the physical situation, but correct distribution functions are used to represent data fluctuations away from the "true values".

Assuming that smaller values of R imply better agreement between the data and model predictions, the definition of p (for continuous distributions of R) is written as:

$$p = \int_{R > R^D} \mathrm{d}R\, P(R|\boldsymbol{\theta}, M)\ . \tag{C.1}$$

The quantity p is the "tail-area" probability to find a result with $R(\boldsymbol{x}) > R^D$, assuming that the model M and the parameters $\boldsymbol{\theta}$ are valid. If the modeling is correct (including that of the data fluctuations), p will have a flat probability distribution between $[0, 1]$. For discrete distributions of R, the integral is replaced by a sum, the p value distribution is no longer continuous, and the cumulative distribution for p will be step-like.

If the existing data are used to modify the parameter values, the extracted p value will be biased to higher values. The amount of bias will depend on many aspects, including the number of data points, the number of parameters, and the priors. We can remove the bias for the number of fitted parameters in χ^2 fits by evaluating the probability of $R = \chi^2$ for $N - n$ degrees-of-freedom, $P(\chi^2|N - n)$, where N is the number of data points and n is the number of parameters fitted [Dri+71], if

- the data fluctuations are Gaussian and independent of the parameters,
- the function to be compared to the data depends linearly on the parameters, and
- the parameters are chosen such that χ^2 is at its global minimum.

In general, the bias introduced by the number of fitted parameters becomes small if $N \gg n$.

p values cannot be turned into probabilistic statements about the model being correct without priors, and statements of "support" for a model directly from the p value behave "incoherently" [Sch96]. Furthermore, approximations used for the distributions of the discrepancy variables, biases introduced when model parameters are fitted and difficulties in extracting reliable information from numerical algorithms used to evaluate the discrepancy variable further complicate their use. p values should therefore be handled with care. Nevertheless, we discuss the use of p values to make judgments about the models at hand, based on a sequence of considerations of the type:

- the p value distribution for a good model is expected to be (reasonably) flat between $[0, 1]$;
- the p values for bad models usually have sharply falling distributions starting at $p = 0$;
- small p values are worrisome; if we know that other models can be reasonably constructed which would have higher p values, then a small p value for the model under consideration indicates that we may have picked a poor model;
- if the p value is not too small, then our model is adequate to describe the existing data.

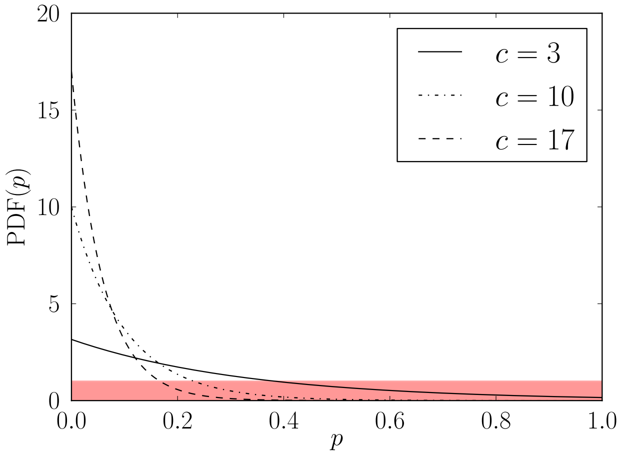


Figure C.1: p value distributions of incorrect models assumed as $P(p|M_i) \approx c_i e^{-c_i p}$. The red area represents the flat distribution under the true model.

C.3 Bayesian motivation

We contend that the use of p values for evaluation of models as just described is essentially Bayesian in character. Following the arguments given above, assume that the p value probability density for a good model, M_0, is flat,

$$P(p|M_0) = 1 \; , \tag{C.2}$$

and that for poor models, $M_i \;\; (i = 1 \ldots n)$, can be represented by

$$P(p|M_i) \approx c_i e^{-c_i p} \; , \tag{C.3}$$

where $c_i \gg 1$ so that the distribution is strongly peaked at 0 and approximately normalized to 1; cf. Fig. C.1. Using Bayes' theorem (2.4), the degree of belief assigned to model M_0 after finding a particular p value is then

$$P(M_0|p) = \frac{P(p|M_0)P(M_0)}{\sum_{i=0}^{n} P(p|M_i)P(M_i)} \; . \tag{C.4}$$

If we take all models to have similar prior degree of belief, then

$$P(M_0|p) \approx \frac{P(p|M_0)}{\sum_{i=0}^{n} P(p|M_i)} \; . \tag{C.5}$$

In the limit $p \to 0$, we have

$$P(M_0|p) \approx \frac{1}{1 + \sum_{i=1}^{n} c_i} \ll 1 \; , \tag{C.6}$$

while for $c_i p \gg 1 \;\; \forall i$

$$P(M_0|p) \approx 1 \; . \tag{C.7}$$

Although this formulation in principle allows for a ranking of models, the vague nature of this procedure indicates that any model which can be constructed to yield a reasonable p value should be retained. A further consideration is that the correct distributions for the data fluctuations are often not known (due to the vague nature of systematic uncertainties) and best guesses are used. This will generally also lead to non-flat p value distributions for good models.

Scientific prejudices (Occam's razor, elegance or aesthetics, etc.) will influence the decision and act as a guide in selecting the "best" model in cases where several good models are available. The preferred quantitative approach for two fully specified models is to use the posterior odds as in (2.14).

A more thorough discussion of goodness of fit, including many detailed examples, is presented in [Bea+11].

D Nonlocal MCMC variants

It is well known that the local random walk MCMC algorithm fails with target densities exhibiting well separated modes. One such example is discussed in Section 3.2. Qualitatively speaking, "well separated" corresponds to many widths of the local proposal function. In general, the larger the dimensionality, the smaller the width to maintain a good acceptance rate. Even if the modes are separated only in one of d dimensions, say x_1, then a local multivariate proposal is tuned in a way that the proposal width in x_1 is roughly $\propto 1/\sqrt{d}$; cf. Algorithm 1. Hence chains are trapped in local modes, and do not mix; they only explore a small subset of the whole parameter space. The chain's current position depends on the starting position, so the asymptotic regime is not reached, and the chain's output is not a sample of the target density, in violation of the basic limit theorem (3.10).

Therefore, it seems reasonable to add nonlocality to the MCMC proposal to cope with multimodality. In this chapter, we sketch some of our early, unsuccessful attempts at such algorithms. Despite considerable effort, we did not arrive at a formulation that is problem-independent or works in $d = 30$, as required for the global fit. The following is shown for completeness, perhaps to warrant further development in this direction. At present, we recommend the procedure based on PMC and outlined in Chapter 4, which is superior and cleaner in every aspect.

D.1 Global-local jumps

In order to overcome the problems with multimodality seen in the example of Section 3.2, we propose to combine global and local jumps. Schematically, the proposal function is divided into two parts

$$q(x|y) = \alpha\, q_{\text{global}}(x) + (1-\alpha) q_{\text{local}}(x|y),\ \alpha \in [0,1]\ . \tag{D.1}$$

The local part is a multivariate normal or student-t, exactly as described in Section 3.1.1. Our primary task is to include as much information about the various regions as possible into the global part to make this approach useful. A second objective is to extract, or learn, the necessary information automatically, without the user having to supply hints manually.

The key realization is that the Markov chains in a adaptive local random walk are good at extracting local information. Now we only have to run multiple chains, and combine the wisdom of the chain ensemble to obtain global information. Our general strategy consists of two phases. First, we perform a prerun with k independent chains doing a local random walk. Next, we extract the information from the combined chain histories to construct q_{global}, and perform the main run.

In each mode, an efficient q_{local} may look very different. To guarantee a good local proposal density, we introduce a discrete hyperparameter h that labels the different modes; e.g., with four modes, $h = 1, 2, 3, 4$. We then perform a random walk on the

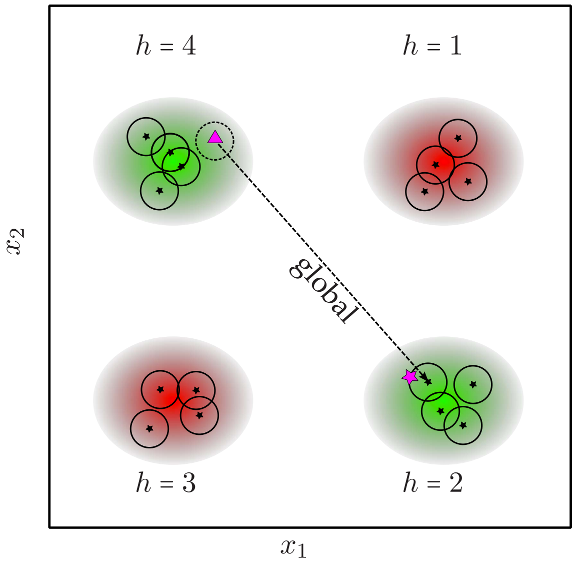


Figure D.1: Schematic of a multimodal bivariate density with four history points in each mode. The solid circles around each history point z (⋆) represent the range of local jumps; a global jump from $h = 4$ (▲) to $h = 2$ (★) is indicated by the dashed arrow.

augmented parameter space of pairs (x, h); assuming local jumps are only within one mode, $q_{\text{local}}(x|y) \to q_{\text{local}}(x|y, h_y)$. As the simplest starting point for q_{global}, we extract representative *history points* from the chains, $\{(z_s, h_s)\}$, and propose new points in their vicinity:

$$q_{\text{global}}(x, h_x) = P(h_x) \sum_s P(z_s) q_{\text{local}}(x|z_s, h_x) \,. \tag{D.2}$$

In words, first select a history point z_s, then draw x locally around z_s. Each history point and mode is weighted by $P(z)$ and $P(h)$ respectively. For efficiency, $P(z)$ should be high when the target density is high. The weight of a mode, $P(h)$, should be proportional to the integral over the mode, which is however not available from the Markov chains. A schematic picture of a global jump is depicted in Fig. D.1.

Several deficiencies of this approach are apparent: the weights $P(z)$ and $P(h)$ are somewhat arbitrary, as is the number of history points and their location. At a more fundamental level, this approach is troublesome when the modes overlap. Because of the hyperparameter, a point x in the overlap appears as two different points, (x, h_1) and (x, h_2). As a consequence, the stationary distribution of the chain is not the target density, because it depends on h.

Refining (D.2), we remove the dependence on h by summing over all history points

$$q_{\text{global}}(x) = \sum_h P(h) \sum_s P(z_s) q_{\text{local}}(x|z_s, h) \,. \tag{D.3}$$

Now h is used only for the local proposal. In a further step, one could remove h entirely by using an identical q_{local} for each mode.

Note the resemblance of q_{global} to a kernel density estimation (KDE) of the target. Many points are needed for a decent interpolation, but this slows down the Metropolis-Hastings steps as a sum over all history points is needed in each iteration to compute the Hastings factor. Furthermore, such an interpolation is (very) poor in $d \gtrsim 3$, another manifestation of the curse of dimensionality. For example, suppose $d = 30$, and the chain has made a few local jumps, now a global jump from x to y is attempted. The coverage being poor, let us assume y is close to only one history point z_y. With a Gaussian local proposal, the typical distance of y to z corresponds to a $\chi^2 = 30$. Assuming x and y far apart, we can neglect the contribution from q_{local}, thus

$$q(y|x) \propto q_{\text{global}}(y|x) \propto \exp(-\,30/2)\ , \tag{D.4}$$

while it takes a few local moves from x to its nearest history point z_x, with a distance of say $\chi^2 = 100$

$$q(x|y) \propto q_{\text{global}}(x|y) \propto \exp(-\,100/2)\ . \tag{D.5}$$

Normalization constants cancel in the Hastings factor, so

$$\frac{q(x|y)}{q(y|x)} \approx \exp(-35)\ . \tag{D.6}$$

The typical change in posterior $P(y)/P(x)$ is much smaller than $\exp(-35)$, hence the Metropolis-Hastings probability of accepting the jump (3.14)

$$\rho(y|x) = \min\left\{\frac{P(y)}{P(x)} \cdot \frac{q(x|y)}{q(y|x)}, 1\right\}\ . \tag{D.7}$$

is dominated by the Hastings factor and largely independent of the target. Essentially all global jumps are rejected; the method fails completely.

Our approach is original work, but in fact strikingly similar to the "bank of clues" by Allanach and Lester [AL08], of which we had not been aware when we investigated the idea. The only key difference is that in [AL08], there is no hyperparameter, and the same kernel is used for every history point. We note that Allanach and Lester [AL08] introduced the idea in particle physics, but they themselves only reinvented what had been known long before in the applied statistics community as "mixture hybrid kernel MCMC" [Tie94]. Allanach and Lester [AL08] remark that the method works only up to $d \approx 10$, in agreement with our numerical tests and the argument leading to (D.6).

D.2 Long jumps

The point of failure of the global-local proposal of Section D.1 is that one needs a very good approximation to the target to avoid the Hastings factor taking over control. This problem cannot occur in the local random walk, because q_{local} is symmetric by construction, thus the Hastings factor is one. The motivation for *long jumps* is to circumvent the need for correct interpolation of the target by a mechanism similar to the local random walk. Suppose there are n local modes $\{x_i^*\}$, then we compute the set of $n(n-1)/2$ translation vectors between the modes. For example, $x_i^* = x_j^* + b(h_i|h_j)$, so $b(h_i|h_j) = -b(h_j|h_i)$. Then the global proposal part is

$$q_{\text{global}}(y|x, h_x) = \sum_{h_y \neq h_x} P(h_y)\delta\left(y - (x + b(h_y|h_x))\right)\ . \tag{D.8}$$

The basic principle is illustrated in Fig. D.2. For x, y in different modes we find

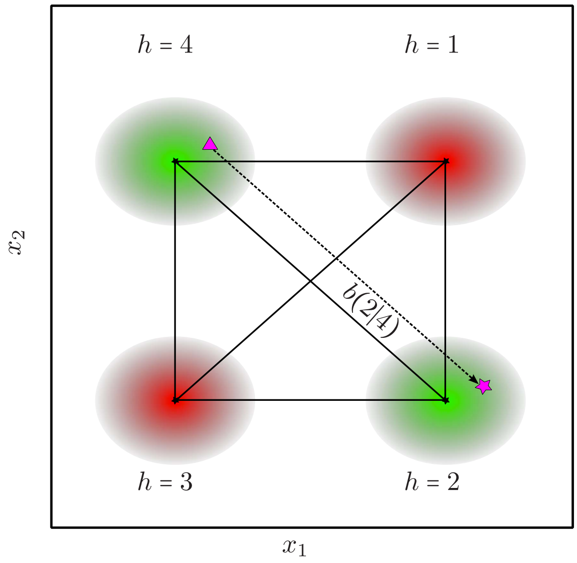


Figure D.2: Schematic of a multimodal bivariate density with the long jumps between modes (⋆) indicated by solid lines. A global jump from the current point (▲) in mode $h = 4$ to the point (★) in mode $h = 2$ is indicated by the dashed arrow that is parallel by construction to the jump connecting the two modes mode 2 and 4.

$q_{\text{local}}(x|y) = q_{\text{local}}(y|x) \approx 0$ and

$$\frac{q(x|y)}{q(y|x)} \approx \frac{P(h_x)}{P(h_y)}\,, \tag{D.9}$$

hence for $P(h) = \text{const}$, acceptance of the new point does not explicitly depend on q, just like in the local symmetric random walk. MCMC with long jumps requires a prerun to adjust the q_{local} before the main run with fixed q is started; mixing due to long jumps works more reliably in higher dimensions and requires less tuning parameters than the global-local approach, but displays its own set of disadvantages. Long jumps work great when individual modes are clones of each other, such that each global jump occurs at a constant level of the target, $P(x) = P(y)$. Then each global jump is accepted with probability $\rho(y|x) = 1$. Difficulties arise with degeneracies, where the local mode is not defined, and with overlapping modes, because the stationary distribution of the Markov chains depends on the hyperparameter, and therefore it is not the target distribution. Marginalizing over h_x in (D.8) using the law of total probability

$$q_{\text{global}}(y|x) = \sum_{h_x} P(h_x|x) \sum_{h_y \neq h_x} P(h_y)\delta\left(y - (x + b(h_y|h_x))\right)\,, \tag{D.10}$$

we require $P(h_x|x)$, the unknown probability that x belongs to mode h_x. It could be defined from a suitable clustering procedure, but this extra complexity would defeat the purpose of a simple proposal with global jumps. For the complicated shapes of the posteriors arising the in the global fit in $d > 20$, the long-jump approach did not work efficiently, and we therefore have not investigated it further.

E The expectation-maximization algorithm

Introduced by Dempster, Laird, and Rubin [DLR77], the expectation-maximization, or EM, algorithm is a powerful tool to solve an optimization problem in the presence of missing or hidden data. EM is applicable when a problem is intractable in terms of the incomplete data, but solvable in closed form with the complete data. Our main application of EM in this work is fitting a mixture density to samples. In that case, the samples are the observed data, the source component of each sample is unknown, and the parameters of interest are the component weights, means, and covariances. First, we discuss the basics of EM along the lines of [Bor04], then treat the mixture fitting example in some detail for illustration in Section E.1.

The EM algorithm is an iterative procedure alternating between the E step and the M step. In the E step, the hidden data are inferred using the current parameter estimates and the observed, incomplete data. Assuming the current estimate of the hidden data, the parameter values are updated. The steps are repeated until a fixed point is reached; in most cases, that is a local maximum, but in pathological cases, EM might converge to a saddle point or even a local minimum.

EM often converges in very few steps, and, as opposed to gradient-based algorithms, copes with very large parameter spaces; e.g., in our global fit, the proposal function requires $\mathcal{O}(50\,000)$ parameters, and the PMC method, a variant of EM, needs about 10 steps to converge.

Let us discuss the foundations of EM in its original context, maximum- likelihood estimation. However, it is important to note that the EM idea is applied in many variants, and not limited to maximizing a likelihood. Suppose D is observed, and we assume D arises from a distribution with parameter $\boldsymbol{\theta}$, then we want to find the mode of

$$L(\boldsymbol{\theta}) \equiv \log P(D|\boldsymbol{\theta}) \; . \tag{E.1}$$

Let Z denote the hidden data, then by the law of total probability

$$P(D|\boldsymbol{\theta}) = \sum_Z P(D|Z,\boldsymbol{\theta}) P(Z|\boldsymbol{\theta}) \; . \tag{E.2}$$

Given the estimate of $\boldsymbol{\theta}$ at time t, $\boldsymbol{\theta}^t$, we need to maximize the difference

$$\begin{aligned} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^t) &= \log\left(\sum_Z P(Z|D,\boldsymbol{\theta}^t)\frac{P(D|Z,\boldsymbol{\theta})P(Z|\boldsymbol{\theta})}{P(Z|D,\boldsymbol{\theta}^t)}\right) - \log P(D|\boldsymbol{\theta}^t) && \text{(E.3)} \\ &\geq \sum_Z P(Z|D,\boldsymbol{\theta}^t)\log\left(\frac{P(D|Z,\boldsymbol{\theta})P(Z|\boldsymbol{\theta})}{P(Z|D,\boldsymbol{\theta}^t)}\right) - \log P(D|\boldsymbol{\theta}^t) && \text{(E.4)} \\ &= \sum_Z P(Z|D,\boldsymbol{\theta}^t)\log\left(\frac{P(D|Z,\boldsymbol{\theta})P(Z|\boldsymbol{\theta})}{P(Z|D,\boldsymbol{\theta}^t)P(D|\boldsymbol{\theta}^t)}\right) && \text{(E.5)} \\ &\equiv \Delta(\boldsymbol{\theta}|\boldsymbol{\theta}^t) \; . && \text{(E.6)} \end{aligned}$$

From (E.3) to (E.4), we used the fact that $\log(\cdot)$ is concave to apply Jensen's inequality,

$$\log \sum_i \alpha_i y_i \geq \sum_i \alpha_i \log(y_i), \quad \sum_i \alpha_i = 1 \; . \tag{E.7}$$

and (E.5) follows from (E.4) because $\sum_Z P(Z|D, \boldsymbol{\theta}^t) = 1$. Defining the function

$$l(\boldsymbol{\theta}|\boldsymbol{\theta}^t) \equiv L(\boldsymbol{\theta}^t) + \Delta(\boldsymbol{\theta}|\boldsymbol{\theta}^t) \; , \tag{E.8}$$

we see that $l(\boldsymbol{\theta}^t|\boldsymbol{\theta}^t) = L(\boldsymbol{\theta}^t)$ due to the definition of conditional probability and that $l(\boldsymbol{\theta}|\boldsymbol{\theta}^t) \leq L(\boldsymbol{\theta})$. Thus by maximizing the approximate likelihood $l(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ with respect to $\boldsymbol{\theta}$ we increase L. Of course, we only gain something if the mode of $l(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ is easier to find than that of $L(\boldsymbol{\theta})$ because of guessing the hidden data; see the example below. For the maximization of (E.8), we can ignore constant terms and focus on the function

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) \equiv \left\{ \sum_Z P(Z|D, \boldsymbol{\theta}^t) \log \left(P(D|Z, \boldsymbol{\theta}) P(Z|\boldsymbol{\theta}) \right) \right\} \tag{E.9}$$

$$= \left\{ \sum_Z P(Z|D, \boldsymbol{\theta}^t) \log P(D, Z|\boldsymbol{\theta}) \right\} \tag{E.10}$$

$$= E_{P(Z|D, \boldsymbol{\theta}^t)} \left[\log P(D, Z|\boldsymbol{\theta}) \right] \; . \tag{E.11}$$

The final result (E.11) explains the terminology: First one has to compute the expectation value under the posterior density of Z conditional on D and $\boldsymbol{\theta}^t$ (E step), then maximize Q to find the next value $\boldsymbol{\theta}^{t+1}$ (M step). In practice, EM is not only applied to maximum likelihood estimation; for example, in the PMC algorithm, one strives to minimize the Kullback-Leibler divergence, so Q is then redefined, but the process of alternating E and M steps to improve the parameter estimate is of general use.

E.1 Gaussian mixture

We want to illustrate the EM algorithm in a simple application. Suppose N samples are drawn independently from a 1D Gaussian mixture density with K components, then the data are $D = \{x^i : i = 1 \ldots N\}$. Note that this example is very closely related to the PMC updates (Section 3.3.1) and the hierarchical clustering (Section 4.2). The incomplete-data likelihood is

$$P(D|\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{j=1}^{K} \alpha_j \mathcal{N}(x^i|\mu_j, \sigma_j) \; , \qquad \sum_j \alpha_j = 1 \; . \tag{E.12}$$

In this case, the hidden data Z comprises the unknown source component of each sample, $Z = \{z^i : i = 1 \ldots N, z^i = 1 \ldots K\}$. Each component is parametrized by $\boldsymbol{\theta}_j = (\alpha_j, \mu_j, \sigma_j)$ for $j = 1 \ldots K$, hence $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

On the one hand, if we knew $\boldsymbol{\theta}$, we could infer the most likely source component Z. On the other hand, if we knew Z, finding the most likely value of $\boldsymbol{\theta}$ would be straightforward, because the problem would factorize into K independent problems, and each one is simple. However, the combined system of equations does not have a closed-form solution, and so we employ the EM algorithm.

Let us fill in the terms to make use of (E.11). The joint distribution of observed and hidden data is

$$\log P(D, Z|\boldsymbol{\theta}) = \sum_i \log\left(P(x^i|z^i)P(z^i)\right) = \log\left(\alpha_{z^i}\mathcal{N}(x^i|\mu_{z^i}, \sigma_{z^i})\right) \,, \tag{E.13}$$

so $\alpha_j = P(z = j)$ is just the prior probability of the j$^{\text{th}}$ component. Thus the probability at time t that x^i was generated from component z^i is just the posterior probability obtained through Bayes' theorem

$$P(z^i|x^i, \boldsymbol{\theta}^t) = \frac{P(x^i|z^i, \boldsymbol{\theta}^t)\alpha^t_{z^i}}{\sum_j P(x^i|j, \boldsymbol{\theta}^t)\alpha^t_{z^i}} = \frac{\alpha^t_{z^i}\mathcal{N}(x^i|\mu^t_{z^i}, \sigma^t_{z^i})}{\sum_j \alpha^t_j \mathcal{N}(x^i|\mu^t_j, \sigma^t_j)} \,, \tag{E.14}$$

where the denominator is the (incomplete-data) likelihood for a single sample $P(x^i|\boldsymbol{\theta})$. Hence $P(Z|D, \boldsymbol{\theta}^t) = \prod_i P(z^i|x^i, \boldsymbol{\theta}^t)$. The evaluation of the expectation value can be done in closed form, and after some algebra one obtains as final result of the E step [Bil98]

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{j=1}^{K}\sum_{i=1}^{N} \log\left(\alpha_j\mathcal{N}(x^i|\mu_j, \sigma_j)\right) P(j|x^i, \boldsymbol{\theta}^t) \,. \tag{E.15}$$

Defining $N_j^{t+1} \equiv \sum_{i=1}^N P(j|x^i, \boldsymbol{\theta}^t)$ as the effective number of samples from component j, the M step leads to the following update of $\boldsymbol{\theta}^{t+1}$:

$$\alpha_j^{t+1} = \frac{N_j^{t+1}}{N} \,, \tag{E.16}$$

$$\mu_j^{t+1} = \frac{\sum_i x^i \, P(j|x^i, \boldsymbol{\theta}^t)}{N_j^{t+1}} \,, \tag{E.17}$$

$$\left(\sigma_j^{t+1}\right)^2 = \frac{\sum_i \left(x^i - \mu_j^{t+1}\right)^2 \, P(j|x^i, \boldsymbol{\theta}^t)}{N_j^{t+1}} \,. \tag{E.18}$$

The results for μ^{t+1} and σ^{t+1} reduce to the well-known sample mean and sample covariance that maximize the likelihood in the case of a single Gaussian, $K = 1$. Note that the E step has been absorbed into the update Equations (E.16) – (E.18), hence due to the simplicity of the problem, there is no need to explicitly calculate the intermediate quantity Q.

F Kernel density estimation

Let us consider kernel density estimation (KDE) [SS05] as an alternative to histograms for the task of *nonparametric* estimation of a probability *density* from a discrete set of samples. The major conceptional difference is that KDE assumes the probability distribution underlying the data is smooth. Whereas a histogram "swallows" samples in discrete bins, KDE approximates the density at a point by a coherent sum over *all* N samples as

$$\hat{P}(\boldsymbol{x}) = \sum_{i=1}^{N} \bar{w}_i K(\boldsymbol{x}|\boldsymbol{x}^i) \ , \tag{F.1}$$

where $\bar{w}_i$ is the self-normalized weight, and $K(\boldsymbol{x}|\boldsymbol{y})$ is a kernel. Many kernels can be used[1], but we restrict ourselves to Gaussian kernels of the form

$$K(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-\frac{1}{h^2}\left(\boldsymbol{x}-\boldsymbol{y}\right)^T\left(\boldsymbol{x}-\boldsymbol{y}\right)\right] , \tag{F.2}$$

because we can then exploit the package FIGTree [Mor+09; Mor10] that provides a speed-up of up to a factor of 100 over a naïve summation through clever caching and ignoring of samples that are too far away to have an impact at a given estimation accuracy. This *fast improved Gauss transform with tree data structure* is the analogy of the fast Fourier transform.

The equivalent of the bin width is the *bandwidth* h for a given kernel. The only free parameter of KDE, h has to be chosen with great care to ensure accurate density estimation [SS05, Ch. 3.2]. Another difference with the histogram approach is the data preprocessing required for KDE in the multivariate case. Its use is quickly seen as follows: suppose we wish to estimate a density in 2D, and the variances are such that $V[x_1] \gg V[x_2]$. Then the euclidean distance between two sample points is dominated by the distance in the first dimension, and the density estimate $\hat{P}$ is nearly independent of the x_2 direction! One way to remedy this situation is to transform the data into almost principal components to arrive at the simple form of (F.2) [SS05, Ch. 3.3]. Another, faster alternative that we follow in this work is to rescale the samples to the unit hypercube; e.g. the first dimension of the j^{th} sample becomes

$$x_1^j \to \frac{x_1^j}{\max_i x_1^i - \min_i x_1^i} \ . \tag{F.3}$$

After rescaling, $h = 0.01$ provides a good starting point for the densities shown in this work. Ultimately, we tune h such that the resulting image is smooth, and features such as the location and diameter of a mode are in good agreement with the histogram results.

From the smoothness assumption, the biggest gain of KDE over the histogram is the smoother output that makes it easier to quickly grasp the important structures of the

[1] The kernel $K(\boldsymbol{x}|\boldsymbol{x}^i) = \mathbf{1}_{x_1, x_1^i \in [a,b]}\left(\boldsymbol{x}|\boldsymbol{x}^i\right)$ reproduces the histogram, but when we use the term KDE, we do *not* include histogram, but only Gaussian kernels.

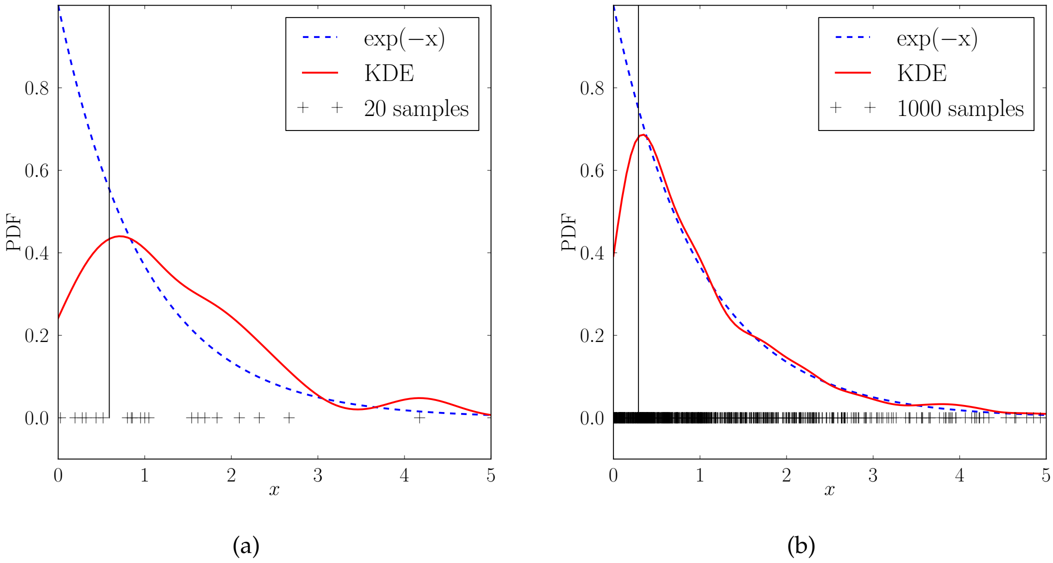


Figure F.1: Illustration of KDE for a set of (a) 20 and (b) 1000 samples from the exponential density. Individual samples are indicated by the "+" symbol. The bandwidth h is chosen according to the conventional rule for samples from the normal density [SS05] depending on the size N and variance σ^2 of the sample as $h = 1.06\,\sigma\,N^{-1/5}$. The density drops at a distance h (vertical black line) away from the minimum $x = 0$ due to the boundary effect.

underlying density. Compare the vertical band present in both Fig. 4.9(b) and Fig. 4.9(c), but in the latter Figure, created with KDE, it is much more pronounced. In addition, KDE is indispensable in computing the $1\,\sigma$ and $2\,\sigma$ contours in the global fit; cf. the Figures in Section 7.2. Consider as a specific example Fig. 7.8(c). Due to the Monte Carlo variance, the histogram approximation yields a $1\,\sigma$ region partitioned into many disjoint regions (not shown), while KDE provides the a-priori more plausible simply connected region shown.

However, there are disadvantages of note, too. First, KDE is more sensitive to a proper bandwidth choice than a histogram is to a proper binning. Second, densities peaking at the boundary of the allowed range are not captured correctly; instead, the KDE interpolation always decreases when within a distance of $\lesssim h$ of the boundary. This *boundary effect* arises because the density is cut off at the boundary, violating the smoothness assumption; it is visible in Fig. 4.9(c) near $(x_1, x_{12}) = (10, \pm 30)$, where the red band does *not* extend to the boundary although the density is constant. To the best of our knowledge, there is a solution to the boundary effect only in $d = 1$ [Jon93], but not for $d \geq 2$.

An illustration of KDE and the boundary effect is displayed in Fig. F.1 for samples from the exponential density. The method produces decent results, even after only 20 samples, but drops for $x \lesssim h$ due to the boundary effect. As more samples are acquired, h is reduced, hence the boundary effect appears in a smaller region. In cases in which the boundary effect is present, we use the histogram approximation instead; cf. the theory predictions in Section 7.6.

Acknowledgments

Three years of Ph.D. have gone by quickly and here is the final word. It is *my* thesis but to say it were all my own work would be a hubris. Many have a share in it. First and foremost, I thank my supervisor Allen Caldwell. His lectures about Bayesian probability and Monte Carlo methods inspired my interest in the whole subject in the first place. A director of the *Max Planck institute for physics* (MPP), he is one busy man. So I am grateful for all the time he spent with me discussing statistical issues. His sound advice kept me on track towards finishing and his support helped me visit a number of great conferences, including the amazing Lindau Nobel laureate meeting 2012.

When it comes to B physics, I am indebted to Christoph Bobeth and Danny van Dyk. It has been a pleasant and fruitful collaboration with them. Christoph's patience in explaining the weird and wonderful ways of EFT and QCD have been of utmost help. Similarly, I greatly enjoyed Danny's mentoring regarding the innards of EOS and fine software engineering in general. I had been programming for many years before meeting Danny but he made me start from scratch again, though at a much *higher* level. Crafting a sampling algorithm to perform the fit — by trial and error as well as long discussions with Allen, Christoph, and Danny — was a very rewarding endeavor.

As a developer of the *Bayesian analysis toolkit* with Dano Kollár and Kevin Kröninger, I learned a lot about developing an actual software product with users, bug fixes, maintenance, releases . . . , and had the opportunity to give tutorials in Rome and Bologna.

Being a student of the *international Max Planck research school* (IMPRS) *on elementary particle physics*, I benefited from the excellent facilities at the MPP and enjoyed the generosity of the German tax payer by visiting conferences near and far, including the workshops at Ringberg castle and Wildbad Kreuth. Many of my IMPRS fellows — a number of them present at the legendary Wolfersdorf summer school in 2008 — provided the necessary distractions to balance long hours in the office. Just to mention a few: Alessandro Manfredini, Daniel Greenwald, Florian Faulstich, Hossein Aghaei-Khozani, Katharina Ecker, Matteo Palermo, Noppadol Mekareeya, Oliver Schlotterer, Peter Patalong, Philipp Weigell, Quaschtl Halter, and Thorsten Rahn. Preparing for the defense, I had the pleasure of learning directly from the experts Dr. Thomas Hahn, Prof. Alois Kabelschacht, and many other junior and senior scientists at the MPP. Many of the above mentioned kindly served as proofreaders to enhance the quality and legibility of this manuscript. Thanks.

Back in highschool times, my teachers Dieter Seckler (†, Hilden) and in particular Steven Stap (South Haven) increased my motivation to pursue studies in physics.

Finally, I received substantial support from outside of academia — a big thanks to my parents and brothers, and especially to my dear girlfriend Ioana Bala-Ciolanescu. I deeply appreciate her help and perseverance during the final weeks of preparing this thesis.

List of abbreviations

CDC central drift chamber

CDF cumulative distribution function

CERN Organisation européenne pour la recherche nucléaire

CKM Cabibbo-Kobayashi-Maskawa

ECAL electromagnetic calorimeter

EFT effective field theory

ESS effective sample size

FCNC flavor changing neutral current

GIM Glashow-Iliopoulos-Maiani

HQET heavy quark effective theory

KDE kernel density estimation

LCDA light cone distribution amplitude

LCSR light cone sum rules

LHC large hadron collider

MAXENT maximum entropy

MCMC Markov chain Monte Carlo

MSSM minimal supersymmetric extension of the SM

NP new physics

OPE operator production expansion

PMC population Monte Carlo

PDF probability density function

QCD quantum chromodynamics

QCDF QCD factorization

RICH ring-imaging Cherenkov

SM standard model

SVD silicon strip vertex detector

TOF time of flight

Bibliography

[Aad+12a] G. Aad et al. [ATLAS Collaboration]. „Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." *Phys.Lett.B* **716** (2012), pp. 1–29. arXiv:1207.7214.

[Aad+12b] G. Aad et al. [ATLAS Collaboration]. „Search for the decay $B_{s,0} \to \mu^+\mu^-$ with the ATLAS detector." *Phys.Lett.B* **713** (2012), pp. 387–407. arXiv:1204.0735.

[Aai+12a] R. Aaij et al. [LHCb Collaboration]. „Determination of the sign of the decay width difference in the B_s system." *Phys.Rev.Lett.* **108** (2012), p. 241801. arXiv:1202.4717.

[Aai+12b] R. Aaij et al. [LHCb collaboration]. „Differential branching fraction and angular analysis of the $B^+ \to K^+\mu^+\mu^-$ decay" (2012). arXiv:1209.4284.

[Aai+12c] R. Aaij et al. [LHCb Collaboration]. „Search for the rare decays $B_s \to \mu^+\mu^-$ and $B_0 \to \mu^+\mu^-$." *Phys.Lett.B* **708** (2012), pp. 55–67. arXiv:1112.1600.

[Aai+12d] R. Aaij et al. [LHCb Collaboration]. „Strong constraints on the rare decays $B_s \to \mu^+\mu^-$ and $B^0 \to \mu^+\mu^-$." *Phys.Rev.Lett.* **108** (2012), p. 231801. arXiv:1203.4493.

[Aal+11a] T. Aaltonen et al. [CDF Collaboration]. „Observation of the Baryonic Flavor-Changing Neutral Current Decay $\Lambda_b \to \Lambda\mu^+\mu^-$." *Phys.Rev.Lett.* **107** (2011), p. 201802. arXiv:1107.3753.

[Aal+11b] T. Aaltonen et al. [CDF Collaboration]. „Search for $B_s \to \mu^+\mu^-$ and $B_d \to \mu^+\mu^-$ Decays with CDF II." *Phys.Rev.Lett.* **107** (2011), p. 239903. arXiv:1107.2304.

[Aal+12] T. Aaltonen et al. [CDF Collaboration]. „Measurements of the Angular Distributions in the Decays $B \to K^{(*)}\mu^+\mu^-$ at CDF." *Phys.Rev.Lett.* **108** (2012), p. 081807. arXiv:1108.0695.

[Aba+02] A. Abashian et al. [Belle Collaboration]. „The Belle Detector." *Nucl. Instrum.Meth.A* **479** (2002), pp. 117–232.

[Aba+10] V. M. Abazov et al. [D0 Collaboration]. „Search for the rare decay $B^0_s \to \mu^+\mu^-$." *Phys.Lett.B* **693** (2010), pp. 539–544. arXiv:1006.3469.

[Aba+95] S. Abachi et al. [D0 Collaboration]. „Search for high mass top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV." *Phys.Rev.Lett.* **74** (1995), pp. 2422–2426. arXiv:hep-ex/9411001.

[Abe+01] K. Abe et al. [Belle Collaboration]. „Observation of large CP violation in the neutral B meson system." *Phys.Rev.Lett.* **87** (2001), p. 091802. arXiv:hep-ex/0107061.

[Abe+95] F. Abe et al. [CDF Collaboration]. „Observation of top quark production in $\bar{p}p$ collisions." *Phys.Rev.Lett.* **74** (1995), pp. 2626–2631. arXiv:hep-ex/9503002.

[Agu04] J. Aguilar-Saavedra. „Top flavor-changing neutral interactions: Theoretical expectations and experimental detection." *Acta Phys.Polon.B* **35** (2004), pp. 2695–2710. arXiv:hep-ph/0409342.

[Aka+03] K. Akai et al. „Commissioning of KEKB." *Nucl.Instrum.Meth.A* **499** (2003), pp. 191–227.

[Aka74] H. Akaike. „A new look at the statistical model identification." *IEEE Transactions on Automatic Control* **19**.6 (1974), pp. 716–723.

[AL08] B. C. Allanach and C. G. Lester. „Sampling using a 'bank' of clues." *Comput.Phys.Commun.* **179** (2008), pp. 256–266. arXiv:0705.0486.

[AlH+10] A. Al-Haydari et al. [QCDSF Collaboration]. „Semileptonic form factors $D \to \pi K$ and $B \to \pi K$ from a fine lattice." *Eur.Phys.J.A* **43** (2010), pp. 107–120. arXiv:0903.1664.

[Ali+00] A. Ali, P. Ball, L. Handoko, and G. Hiller. „A Comparative study of the decays $B \to (K, K^{*)}\ell^+\ell^-$ in standard model and supersymmetric theories." *Phys.Rev.D* **61** (2000), p. 074024. arXiv:hep-ph/9910221.

[Alo+11] A. K. Alok et al. „New Physics in $b \to s\mu^+\mu^-$: CP-Conserving Observables." *JHEP* **1111** (2011), p. 121. arXiv:1008.2367.

[Alt+09] W. Altmannshofer et al. „Symmetries and Asymmetries of $B \to K^*\mu^+\mu^-$ Decays in the Standard Model and Beyond." *JHEP* **0901** (2009), p. 019. arXiv:0811.1214.

[APS12] W. Altmannshofer, P. Paradisi, and D. M. Straub. „Model-Independent Constraints on New Physics in $b \to s$ Transitions." *JHEP* **1204** (2012), p. 008. arXiv:1111.1257.

[AS12] W. Altmannshofer and D. M. Straub. „Cornering New Physics in $b \to s$ Transitions." *JHEP* **1208** (2012), p. 121. arXiv:1206.0273.

[ATL12] ATLAS, CMS, and LHCb Collaborations. „Search for the rare decays $B^0_{(s)} \to \mu^+\mu^-$ at the LHC with the ATLAS, CMS and LHCb experiments" (2012). LHCb-CONF-2012-017.

[ATR06] R. R. de Austri, R. Trotta, and L. Roszkowski. „A Markov chain Monte Carlo analysis of the CMSSM." *JHEP* **0605** (2006), p. 002. arXiv:hep-ph/0602028.

[Aub+01] B. Aubert et al. [BABAR Collaboration]. „Observation of CP violation in the B^0 meson system." *Phys.Rev.Lett.* **87** (2001), p. 091801. arXiv:hep-ex/0107013.

[Aub+02] B. Aubert et al. [BABAR Collaboration]. „The BaBar detector." *Nucl. Instrum.Meth.A* **479** (2002), pp. 1–116. arXiv:hep-ex/0105044.

[Aub+08] B. Aubert et al. [BABAR Collaboration]. „Measurement of Time-Dependent CP Asymmetry in $B^0 \to K^0_S\pi^0\gamma$ Decays." *Phys.Rev.D* **78** (2008), p. 071102. arXiv:0807.3103.

[Aub+09] B. Aubert et al. [BABAR Collaboration]. „Measurement of Branching Fractions and CP and Isospin Asymmetries in $B \to K^*(892)\gamma$ Decays.“ *Phys. Rev. Lett.* **103** (2009), p. 211802. arXiv:0906.2177.

[Aus12] Australian Institute for High Energy Physics. 2012. http://aushep.org.au/resources.html.

[Baz+12] A. Bazavov et al. [Fermilab Lattice and MILC Collaborations]. „B- and D-meson decay constants from three-flavor lattice QCD.“ *Phys.Rev.D* **85** (2012), p. 114506. arXiv:1112.3051.

[BB00] M. J. Bayarri and J. O. Berger. „P Values for Composite Null Models.“ *J.Am.Stat.Assoc* **95** (2000), pp. 1127–1142.

[BB58] G. Barnard and T. Bayes. „Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances.“ *Biometrika* **45** (1958), pp. 293–315.

[BB92] J. O. Berger and J. M. Bernardo. „On the development of reference priors.“ *Bayesian statistics* **4** (1992), pp. 35–60.

[BBF11] M. Beylich, G. Buchalla, and T. Feldmann. „Theory of $B \to K^*\ell^+\ell^-$ decays at high q^2: OPE and quark-hadron duality.“ *Eur.Phys.J.C* **71** (2011), p. 1635. arXiv:1101.5118.

[BBL06] P. Ball, V. Braun, and A. Lenz. „Higher-twist distribution amplitudes of the K meson in QCD.“ *JHEP* **0605** (2006), p. 004. arXiv:hep-ph/0603063.

[BBL96] G. Buchalla, A. J. Buras, and M. E. Lautenbacher. „Weak decays beyond leading logarithms.“ *Rev.Mod.Phys.* **68** (1996), pp. 1125–1144. arXiv:hep-ph/9512380.

[BC11] F. Beaujean and A. Caldwell. „A test statistic for weighted runs.“ *J.Statist. Plann.Inference* **141** (2011), pp. 3437–3446. arXiv:1005.3233.

[BCL09] C. Bourrely, I. Caprini, and L. Lellouch. „Model-independent description of $B \to \pi\ell\nu$ decays and a determination of $|V_{ub}|$.“ *Phys.Rev.D* **79** (2009), p. 013008. arXiv:0807.2722.

[Bea09] F. Beaujean. *Monte Carlo Methods and Bayesian Data Analysis*. Lecture held at Universidad de Costa Rica. 2009.

[Bea+11] F. Beaujean, A. Caldwell, D. Kollár, and K. Kröninger. „p-values for model evaluation.“ *Phys.Rev.D* **83** (2011), p. 012004. arXiv:1011.1674.

[Bea+12] F. Beaujean, C. Bobeth, D. van Dyk, and C. Wacker. „Bayesian Fit of Exclusive $b \to s\bar{\ell}\ell$ Decays: The Standard Model Operator Basis.“ *JHEP* **1208** (2012), p. 030. arXiv:1205.1838.

[Bec+12] D. Becirevic, E. Kou, A. Le Yaouanc, and A. Tayduganov. „Future prospects for the determination of the Wilson coefficient $C'_{7\gamma}$.“ *JHEP* **1208** (2012), p. 090. arXiv:1206.1502.

[Beh+12] A. Behring, C. Gross, G. Hiller, and S. Schacht. „Squark Flavor Implications from $B \to K^*\ell^+\ell^-$.“ *JHEP* **1208** (2012), p. 152. arXiv:1205.1500.

[Ber05] J. M. Bernardo. *Reference analysis*. Handbook of Statistics 25. Amsterdam: Elsevier, 2005.

[Ber+12] J. Beringer et al. [Particle Data Group]. „Review of Particle Physics.“ *Phys. Rev. D* **86** (2012), p. 010001.

[Ber79] J. M. Bernardo. „Reference Posterior Distributions for Bayesian Inference.“ *J.R.Stat.Soc.Series.B Stat. Methodol* **41** (1979), pp. 113–147.

[BF01] M. Beneke and T. Feldmann. „Symmetry breaking corrections to heavy to light B meson form-factors at large recoil.“ *Nucl.Phys.B* **592** (2001), pp. 3–34. arXiv:hep-ph/0008255.

[BFS01] M. Beneke, T. Feldmann, and D. Seidel. „Systematic approach to exclusive $B \to V\ell^+\ell^-, V\gamma$ decays.“ *Nucl.Phys.B* **612** (2001), pp. 25–58. arXiv:hep-ph/0106067.

[BFS05] M. Beneke, T. Feldmann, and D. Seidel. „Exclusive radiative and electroweak $b \to d$ and $b \to s$ penguin decays at NLO.“ *Eur.Phys.J.C* **41** (2005), pp. 173–188. arXiv:hep-ph/0412400.

[BFW10] A. Bharucha, T. Feldmann, and M. Wick. „Theoretical and Phenomenological Constraints on Form Factors for Radiative and Semi-Leptonic B-Meson Decays.“ *JHEP* **1009** (2010), p. 090. arXiv:1004.3249.

[BGH00] A. Buras, P. Gambino, and U. Haisch. „Electroweak penguin contributions to nonleptonic $\Delta F = 1$ decays at NNLO.“ *Nucl.Phys.B* **570** (2000), pp. 117–154. arXiv:hep-ph/9911250.

[BGP10] P. Bruneau, M. Gelgon, and F. Picarougne. „Parsimonious reduction of Gaussian mixture models with a variational-Bayes approach.“ *Patt.Recog.* **43**.3 (2010), pp. 850–858.

[BHD10] C. Bobeth, G. Hiller, and D. van Dyk. „The Benefits of $\bar{B} \to \bar{K}^* l^+ l^-$ Decays at Low Recoil.“ *JHEP* **1007** (2010), p. 098. arXiv:1006.5013.

[BHD11a] C. Bobeth, G. Hiller, and D. van Dyk. „Angular analysis of $B \to V(\to P_1 P_2) l^+ l^-$ decays.“ *J.Phys.Conf.Ser.* **335** (2011), p. 012038. arXiv:1105.2659.

[BHD11b] C. Bobeth, G. Hiller, and D. van Dyk. „More Benefits of Semileptonic Rare B Decays at Low Recoil: CP Violation.“ *JHEP* **1107** (2011), p. 067. arXiv:1105.0376.

[BHP07] C. Bobeth, G. Hiller, and G. Piranishvili. „Angular distributions of $\bar{B} \to \bar{K}\ell^-\ell^+$ decays.“ *JHEP* **0712** (2007), p. 040. arXiv:0709.4174.

[BHP08] C. Bobeth, G. Hiller, and G. Piranishvili. „CP Asymmetries in bar $\bar{B} \to \bar{K}^*(\to \bar{K}\pi)\bar{\ell}\ell$ and Untagged $\bar{B}_s$, $B_s \to \phi(\to K^+K^-)\bar{\ell}\ell$ Decays at NLO.“ *JHEP* **0807** (2008), p. 106. arXiv:0805.2525.

[Bia+10] M. Biagini et al. [SuperB Collaboration]. „SuperB Progress Reports: The Collider“ (2010). arXiv:1009.6178.

[Bil98] J. Bilmes. „A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models“ (1998). http://www.cns.nyu.edu/~lcv/meeting/summ06/papers/em_bilmes.pdf.

[BK00] K. Babu and C. F. Kolda. „Higgs mediated $B^0 \to \mu^+\mu^-$ in minimal supersymmetry.“ *Phys.Rev.Lett.* **84** (2000), pp. 228–231. arXiv:hep-ph/9909476.

[BLM07] D. Bećirević, V. Lubicz, and F. Mescia. „An estimate of the $B \to K^*\gamma$ decay form factor.“ *Nucl.Phys.B* **769** (2007), pp. 31–43. arXiv:hep-ph/0611295.

[Blo+06] C. Blocker et al. „Simple Facts about P-Values“ (2006). CDF-MEMO-8023.

[BMU00] C. Bobeth, M. Misiak, and J. Urban. „Photonic penguins at two loops and $m(t)$ dependence of $\mathcal{B}(B \to X(s)\ell^+\ell^-)$.“ *Nucl.Phys.B* **574** (2000), pp. 291–330. arXiv:hep-ph/9910220.

[Bob+01] C. Bobeth, T. Ewerth, F. Kruger, and J. Urban. „Analysis of neutral Higgs boson contributions to the decays $\bar{B}_s \to \ell^+\ell^-$ and $\bar{B} \to K\ell^+\ell^-$.“ *Phys.Rev.D* **64** (2001), p. 074014. arXiv:hep-ph/0104284.

[Bob+04] C. Bobeth, P. Gambino, M. Gorbahn, and U. Haisch. „Complete NNLO QCD analysis of $\bar{B} \to X(s)\ell^+\ell^-$ and higher order electroweak effects.“ *JHEP* **0404** (2004), p. 071. arXiv:hep-ph/0312090.

[Bob+12] C. Bobeth, G. Hiller, D. van Dyk, and C. Wacker. „The Decay $B \to K\ell^+\ell^-$ at Low Hadronic Recoil and Model-Independent $\Delta B = 1$ Constraints.“ *JHEP* **1201** (2012), p. 107. arXiv:1111.2558.

[Bon+06] M. Bona et al. [UTfit Collaboration]. „The Unitarity Triangle Fit in the Standard Model and Hadronic Parameters from Lattice QCD: A Reappraisal after the Measurements of $\Delta m(s)$ and $\mathcal{B}(B \to \tau\nu(\tau))$.“ *JHEP* **0610** (2006), p. 081. arXiv:hep-ph/0606167. We use the results of the 2010 fit from http://www.utfit.org.

[Bor04] S. Borman. „The Expectation Maximization Algorithm – A short tutorial.“ http://www.seanborman.com/publications/EM_algorithm.pdf. 2004.

[BR10] A. Bharucha and W. Reece. „Constraining new physics with $B \to K^*\mu^+\mu^-$ in the early LHC era.“ *Eur.Phys.J.C* **69** (2010), pp. 623–640. arXiv:1002.4310.

[Bru+12] K. de Bruyn et al. „Probing New Physics via the $B_s^0 \to \mu^+\mu^-$ Effective Lifetime.“ *Phys.Rev.Lett.* **109** (2012), p. 041801. arXiv:1204.1737.

[BS12] D. Bećirević and E. Schneider. „On transverse asymmetries in $B \to K^*\ell^+\ell^-$.“ *Nucl.Phys.B* **854** (2012), pp. 321–339. arXiv:1106.3283.

[BZ05a] P. Ball and R. Zwicky. „$B_{d,s} \to \rho, \omega, K^*, \phi$ Decay Form Factors from Light-Cone Sum Rules Revisited.“ *Phys.Rev.D* **71** (2005), p. 014029. arXiv:hep-ph/0412079.

[BZ05b] P. Ball and R. Zwicky. „New results on $B \to \pi$, K, η decay form factors from light-cone sum rules.“ *Phys.Rev.D* **71** (2005), p. 014015. arXiv:hep-ph/0406232.

[Cab63] N. Cabibbo. „Unitary Symmetry and Leptonic Decays.“ *Phys.Rev.Lett.* **10** (1963), pp. 531–533.

[Cal10] A. Caldwell. *Data Analysis and Monte Carlo Methods*. Lecture held at Technische Universität München. 2010. http://www.mpp.mpg.de/~caldwell/ss10.html.

[Cal11] A. Caldwell. „Signal discovery in sparse spectra: a Bayesian analysis.“ In proceedings of: PHYSTAT2011, CERN-2011-006. 2011, pp. 138–142.

[Cap+04] O. Cappé, A. Guillin, J. Marin, and C. Robert. „Population Monte Carlo.“ *J.Comput.Graph.Statist.* **13**.4 (2004), pp. 907–929.

[Cap+08] O. Cappé et al. „Adaptive importance sampling in general mixture classes.“ *Stat.Comp.* **18** (2008), pp. 447–459.

[Cha+05] J. Charles et al. [CKMfitter Group]. „CP violation and the CKM matrix: Assessing the impact of the asymmetric B factories.“ *Eur.Phys.J.C* **41** (2005), pp. 1–131. arXiv:hep-ph/0406184.

[Cha+11] S. Chatrchyan et al. [CMS Collaboration]. „Search for B_s and $B \to$ dimuon decays in pp collisions at 7 TeV.“ *Phys.Rev.Lett.* **107** (2011), p. 191802. arXiv:1107.5834.

[Cha+12a] S. Chatrchyan et al. [CMS Collaboration]. „Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC.“ *Phys.Lett.B* **716** (2012), pp. 30–61. arXiv:1207.7235.

[Cha+12b] S. Chatrchyan et al. [CMS Collaboration]. „Search for $B_s^0 \to \mu^+\mu^-$ and $B^0 \to \mu^+\mu^-$ decays.“ *JHEP* **1204** (2012), p. 033. arXiv:1203.3976.

[Cha+99] J. Charles et al. „Heavy to light form-factors in the heavy mass to large energy limit of QCD.“ *Phys.Rev.D* **60** (1999), p. 014001. arXiv:hep-ph/9812358.

[CHM07] M. Czakon, U. Haisch, and M. Misiak. „Four-Loop Anomalous Dimensions for Radiative Flavour-Changing Decays.“ *JHEP* **0703** (2007), p. 008. arXiv:hep-ph/0612329.

[CJ01] S. Chib and I. Jeliazkov. „Marginal likelihood from the Metropolis-Hastings output.“ *J.Amer.Statist.Assoc.* **96**.453 (2001), pp. 270–281.

[Cla12] P. Clarke [LHCb Collaboration]. „Tagged time-dependent angular analysis of $B_s^0 \to J/\psi\phi$ decays at LHCb.“ In proceedings of: 47th Rencontres de Moriond: Electroweak Interactions and Unified Theories, LHCb-CONF-2012-002. 2012.

[Cle+98] B. Cleveland et al. „Measurement of the solar electron neutrino flux with the Homestake chlorine detector.“ *Astrophys.J.* **496** (1998), pp. 505–526.

[CMM97] K. G. Chetyrkin, M. Misiak, and M. Munz. „Weak radiative B meson decay beyond leading logarithms.“ *Phys.Lett.B* **400** (1997), pp. 206–219. arXiv:hep-ph/9612313.

[Coa+00] T. Coan et al. [CLEO Collaboration]. „Study of exclusive radiative B meson decays.“ *Phys.Rev.Lett.* **84** (2000), pp. 5283–5287. arXiv:hep-ex/9912057.

[Cor+12] J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert. „Adaptive Multiple Importance Sampling.“ *Scand.J.Stat.* (2012).

[Cor12] M. D. Corcoran [CDF and D0 Collaborations]. „CP violation and rare B_s decays at the Tevatron.“ *Nuovo Cim.* **C035N1** (2012), pp. 273–280.

[Cra99] H. Cramér. *Mathematical methods of statistics*. Princeton University Press, 1999.

[Cro10] G. Crooks. „The Amoroso Distribution“ (2010). arXiv:1005.3274.

[DAg03] G. D'Agostini. *Bayesian Reasoning in Data Analysis: A Critical Introduction*. World Scientific, 2003.

[DE10] C. Donnelly and P. Embrechts. „The devil is in the tails: actuarial mathematics and the subprime mortgage crisis.“ *Astin Bulletin* **40**.1 (2010), pp. 1–33.

[Des+11] S. Descotes-Genon, D. Ghosh, J. Matias, and M. Ramon. „Exploring New Physics in the $\mathcal{C}_7 - \mathcal{C}_7'$ plane.“ *JHEP* **1106** (2011), p. 099. arXiv:1104.3342.

[DLR77] A. Dempster, N. Laird, and D. Rubin. „Maximum likelihood from incomplete data via the EM algorithm.“ *J.R.Stat.Soc.Ser.B Stat. Methodol.* **39**.1 (1977), pp. 1–38.

[Dri+71] D. Drijard et al. *Statistical methods in experimental physics*. North-Holland, 1971.

[Dyk+12] D. van Dyk, F. Beaujean, C. Bobeth, and C. Wacker [EOS Collaboration] (2012). http://project.het.physik.tu-dortmund.de/eos/.

[Dyk12] D. van Dyk. „The Decays $\bar{B} \to \bar{K}^{(*)}\ell^+\ell^-$ at Low Recoil and their Constraints on New Physics.“ PhD thesis, Technical University of Dortmund, 2012. HDL: 2003/29514.

[EAPG09] A. El Attar, A. Pigeau, and M. Gelgon. „Fast aggregation of Student mixture models.“ In proceedings of: European Signal Processing Conference (Eusipco 2009). 2009.

[EB64] F. Englert and R. Brout. „Broken Symmetry and the Mass of Gauge Vector Mesons.“ *Phys.Rev.Lett.* **13** (1964), pp. 321–323.

[Ege+08] U. Egede et al. „New observables in the decay mode $\bar{B}_d \to \bar{K}^{*0}\ell^+\ell^-$.“ *JHEP* **0811** (2008), p. 032. arXiv:0807.2589.

[Ege+10] U. Egede et al. „New physics reach of the decay mode $\bar{B} \to \bar{K}^{*0}\ell^+\ell^-$.“ *JHEP* **1010** (2010), p. 056. arXiv:1005.0571.

[Fer34] E. Fermi. „Versuch einer Theorie der β-Strahlen. I.“ *Zeitschrift für Physik* **88** (1934), pp. 161–177.

[Fey42] R. Feynman. *Feynman's thesis: a new approach to quantum theory*. Ed. by L. Brown. World Scientific, 1942.

[FHB09] F. Feroz, M. Hobson, and M. Bridges. „MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics.“ *Mon. Not.Roy.Astron.Soc.* **398** (2009), pp. 1601–1614. arXiv:0809.3437.

[Fla+09] H. Flacher et al. „Revisiting the Global Electroweak Fit of the Standard Model and Beyond with Gfitter.“ *Eur.Phys.J.C* **60** (2009), pp. 543–583. arXiv:0811.0009.

[FM03] T. Feldmann and J. Matias. „Forward backward and isospin asymmetry for $B \to K^*\ell^+\ell^-$ decay in the standard model and in supersymmetry.“ *JHEP* **0301** (2003), p. 074. arXiv:hep-ph/0212158.

[Fuj09] M. Fujikawa. „Measurement of Branching Fraction and Time-dependent CP Asymmetry Parameters in $B^0 \to K^0\pi^0$ Decays.“ PhD thesis, Nara Women's University, 2009.

[Gew89] J. Geweke. „Bayesian inference in econometric models using Monte Carlo integration.“ *Econometrica* **57**.6 (1989), pp. 1317–1339.

[GH01] P. Gambino and U. Haisch. „Complete electroweak matching for radiative B decays.“ *JHEP* **0110** (2001), p. 020. arXiv:hep-ph/0109058.

[GH05] M. Gorbahn and U. Haisch. „Effective Hamiltonian for non-leptonic $|\Delta F| = 1$ decays at NNLO in QCD.“ *Nucl.Phys.B* **713** (2005), pp. 291–332. arXiv:hep-ph/0411071.

[GHK64] G. Guralnik, C. Hagen, and T. Kibble. „Global Conservation Laws and Massless Particles.“ *Phys.Rev.Lett.* **13** (1964), pp. 585–587.

[GHM05] M. Gorbahn, U. Haisch, and M. Misiak. „Three-loop mixing of dipole operators.“ *Phys.Rev.Lett.* **95** (2005), p. 102004. arXiv:hep-ph/0504194.

[GIM70] S. Glashow, J. Iliopoulos, and L. Maiani. „Weak Interactions with Lepton-Hadron Symmetry.“ *Phys.Rev.D* **2** (1970), pp. 1285–1292.

[GM98] A. Gelman and X.-L. Meng. „Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling.“ *Stat.Sci.* **13**.2 (1998), pp. 163–185.

[GP04] B. Grinstein and D. Pirjol. „Exclusive rare $B \to K^* \ell^+ \ell^-$ - decays at low recoil: Controlling the long-distance effects.“ *Phys.Rev.D* **70** (2004), p. 114005. arXiv:hep-ph/0404250.

[GR04] J. Goldberger and S. Roweis. „Hierarchical clustering of a mixture model.“ *Adv.Neur.Info.Proc.Syst.* **17** (2004), p. 505.

[GR92] A. Gelman and D. Rubin. „Inference from iterative simulation using multiple sequences.“ *Stat.Sci.* **7**.4 (1992), pp. 457–472.

[Har83] J. A. Hartigan. *Bayes theory*. Springer, 1983.

[Has70] W. Hastings. „Monte Carlo sampling methods using Markov chains and their applications.“ *Biometrika* **57**.1 (1970), pp. 97–109.

[Hei05] J. Heinrich. „Bayesian limit software: multi-channel with correlated backgrounds and efficiencies“ (2005). CDF-MEMO-7587.

[Her+77] S. Herb et al. „Observation of a Dimuon Resonance at 9.5-GeV in 400-GeV Proton-Nucleus Collisions.“ *Phys.Rev.Lett.* **39** (1977), pp. 252–255.

[HH12] C. Hambrock and G. Hiller. „Extracting $B \to K^*$ Form Factors from Data“ (2012). arXiv:1204.4444.

[Hig64] P. W. Higgs. „Broken Symmetries and the Masses of Gauge Bosons.“ *Phys.Rev.Lett.* **13** (1964), pp. 508–509.

[HOVD11] L. Hoogerheide, A. Opschoor, and H. Van Dijk. „A Class of Adaptive EM-Based Importance Sampling Algorithms for Efficient and Robust Posterior and Predictive Simulation.“ *Tinbergen Institute Discussion Paper* **004** (2011).

[HST01] H. Haario, E. Saksman, and J. Tamminen. „An Adaptive Metropolis Algorithm.“ *Bernoulli* **7**.2 (2001), pp. 223–242.

[Hub+06] T. Huber, E. Lunghi, M. Misiak, and D. Wyler. „Electromagnetic logarithms in $\bar{B} \to X(s)\ell^+\ell^-$.“ *Nucl.Phys.B* **740** (2006), pp. 105–137. arXiv:hep-ph/0512066.

[Hut12] D. Hutchcroft [LHCb Collaboration]. „Rare decays at LHCb." In proceedings of: BEACH 2012, Wichita, KS, USA. 2012.

[IW90] N. Isgur and M. B. Wise. „Relationship between form-factors in semileptonic $\bar{b}$ and D decays and exclusive rare $\bar{b}$ decays." *Phys.Rev.D* **42** (1990), pp. 2388–2391.

[Jam75] F. James. „MINUIT — a system for function minimization and analysis of the parameter errors and correlations." *Comput.Phys.Commun.* **10** (1975), pp. 343–367.

[JB03] E. T. Jaynes and G. L. Bretthorst. *Probability theory*. Cambridge University Press, 2003.

[Jef39] H. Jeffreys. *Theory of Probability*. Oxford: Clarendon Press, 1939.

[Jon93] M. C. Jones. „Simple boundary correction for kernel density estimation." *J.Stat.Comp.* **3** (1993), pp. 135–146.

[Ken+04] M. G. Kendall et al. *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Arnold, 2004.

[Kho+10] A. Khodjamirian, T. Mannel, A. Pivovarov, and Y.-M. Wang. „Charm-loop effect in $B \to K^{(*)}\ell^+\ell^-$ and $B \to K^*\gamma$." *JHEP* **1009** (2010), p. 089. arXiv:1006.4945.

[Kil+10] M. Kilbinger et al. „Bayesian model comparison in cosmology with Population Monte Carlo." *Mon.Not.R.Astron.Soc* **405**.4 (2010), pp. 2381–2390. arXiv:0912.1614.

[Kil+11] M. Kilbinger et al. *PMC lib v1.0*. 2011. http://www2.iap.fr/users/kilbinge/CosmoPMC/.

[KL51] S. Kullback and R. Leibler. „On information and sufficiency." *Ann.Math. Stat.* **22**.1 (1951), pp. 79–86.

[KM05] F. Kruger and J. Matias. „Probing new physics via the transverse amplitudes of $B_0 \to K^{*0}(\to K^-\pi^+)\ell^+\ell^-$ at large recoil." *Phys.Rev.D* **71** (2005), p. 094009. arXiv:hep-ph/0502060.

[KM73] M. Kobayashi and T. Maskawa. „CP Violation in the Renormalizable Theory of Weak Interaction." *Prog.Theor.Phys.* **49** (1973), pp. 652–657.

[Kod+01] K. Kodama et al. [DONUT Collaboration]. „Observation of tau neutrino interactions." *Phys.Lett.B* **504** (2001), pp. 218–224. arXiv:hep-ex/0012035.

[Kol33] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933.

[Lap20] P. S. de Laplace. *Théorie analytique des probabilités*. Paris, 1820.

[LC95] J. Liu and R. Chen. „Blind deconvolution via sequential imputations." *J.Amer.Statist.Assoc.* **90**.430 (1995), pp. 567–576.

[Lee+12] J. Lees et al. [BABAR Collaboration]. „Measurement of Branching Fractions and Rate Asymmetries in the Rare Decays $B \to K^{(*)}l^+l^-$." *Phys.Rev.D* **86** (2012), p. 032012. arXiv:1204.3933.

[LHC08] LHCb Collaboration. „The LHCb Detector at the LHC.“ *J.Inst.* **3** (2008), S08005.

[LHC98] LHCb Collaboration. *LHCb : Technical Proposal*. Geneva: CERN, 1998.

[Liu+09] Z. Liu et al. „Form factors for rare B decays: Strategy, methodology, and numerical study.“ *PoS* **LAT2009** (2009). arXiv:`0911.2370`.

[Liu+11] Z. Liu et al. „A Lattice calculation of $B \to K^{(*)}\ell^+\ell^-$ form factors“ (2011). arXiv:`1101.2726`.

[LLVdW10] J. Laiho, E. Lunghi, and R. S. Van de Water. „Lattice QCD inputs to the CKM unitarity triangle analysis.“ *Phys.Rev.D* **81** (2010), p. 034503. arXiv:`0910.2928`.

[Mar06] A. Markov. „Extension of the law of large numbers to dependent events.“ *Bull. Soc. Phys. Math. Kazan* **2**.15 (1906). In Russian, pp. 155–156.

[Mat+12] J. Matias, F. Mescia, M. Ramon, and J. Virto. „Complete Anatomy of $\bar{B}_d \to \bar{K}^{*0}(\to K\pi)\ell^+\ell^-$ and its angular distribution.“ *JHEP* **1204** (2012), p. 104. arXiv:`1202.4266`.

[McN+12] C. McNeile et al. „High-Precision f_{B_s} and HQET from Relativistic Lattice QCD.“ *Phys.Rev.D* **85** (2012), p. 031503. arXiv:`1110.4510`.

[Met+53] N. Metropolis et al. „Equation of state calculations by fast computing machines.“ *J.Chem.Phys.* **21** (1953), p. 1087.

[Mor+09] V. I. Morariu et al. „Automatic online tuning for fast Gaussian summation.“ *Adv.Neural.Info.Proc.Syst* **21** (2009), pp. 1113–1120.

[Mor10] V. Morariu. *FIGTree v0.9.3*. 2010.
`http://www.umiacs.umd.edu/~morariu/figtree/`.

[MS04] M. Misiak and M. Steinhauser. „Three loop matching of the dipole operators for $b \to s\gamma$ and $b \to sg$.“ *Nucl.Phys.B* **683** (2004), pp. 277–305. arXiv:`hep-ph/0401041`.

[MXZ08] F. Muheim, Y. Xie, and R. Zwicky. „Exploiting the width difference in $B_s \to \phi\gamma$.“ *Phys.Lett.B* **664** (2008), pp. 174–179. arXiv:`0802.0876`.

[Nak+04] M. Nakao et al. [Belle Collaboration]. „Measurement of the $B \to K^*(892)\gamma$ branching fractions and asymmetries.“ *Phys.Rev.D* **69** (2004), p. 112001. arXiv:`hep-ex/0402042`.

[Nak+10] K. Nakamura et al. [Particle Data Group]. „Review of particle physics.“ *J.Phys.G* **37** (2010), p. 075021.

[Nat+06] Z. Natkaniec et al. „Status of the Belle silicon vertex detector.“ *Nucl. Instrum.Meth.A* **560** (2006), pp. 1–4.

[NMU51] J. von Neumann, N. Metropolis, and S. Ulam. „Monte Carlo Method.“ *National Bureau of Standards/Applied Math. Series* **12** (1951), pp. 36–38.

[Nor12] F. Norrod. „Perspective Across The Technology Landscape.“ In proceedings of: CHEP 2012, New York, NY, USA. 2012.

[Occ95] W. of Occam. *Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi*. London, 1495.

[OZ00] A. Owen and Y. Zhou. „Safe and effective importance sampling.“ *J.Amer. Statist.Assoc.* **95**.449 (2000), pp. 135–143.

[Par12] C. Parkinson [LHCb Collaboration]. „Differential branching fraction and angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay.“ In proceedings of: 47th Rencontres de Moriond: QCD Sessions, `LHCb-CONF-2012-008`. 2012.

[Poi12] V. Poireau [BaBar Collaboration]. „A selection of recent results from the BaBar experiment“ (2012). arXiv:`1205.2201`.

[RC04] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.

[Rea02] A. L. Read. „Presentation of search results: The CL(s) technique.“ *J.Phys.G* **28** (2002), pp. 2693–2704.

[Ree10] W. R. Reece. „Exploiting angular correlations in the rare decay $B \to K^*\mu^+\mu^-$ at LHCb.“ `CERN-THESIS-2010-095`. PhD thesis, Imperial College London, 2010.

[RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. „Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms.“ *Ann.Appl.Probab.* **7**.1 (1997), pp. 110–120.

[SBB01] T. Sellke, M. J. Bayarri, and J. O. Berger. „Calibration of p Values for Testing Precise Null Hypotheses.“ *Am.Stat* **55** (2001), pp. 62–71.

[Sch78] G. Schwarz. „Estimating the dimension of a model.“ *Ann.Stat.* **6**.2 (1978), pp. 461–464.

[Sch96] M. J. Schervish. „P Values: What They Are and What They Are Not.“ *Am.Stat* **50** (1996), pp. 203–206.

[Sha48] C. E. Shannon. „The Mathematical Theory of Communication.“ *Bell System Techn.J.* **27** (1948), pp. 379–423.

[Shi11] K. Shibata. „Status and schedule of SuperKEKB.“ *PoS* **EPS-HEP2011** (2011), p. 38.

[Sim+10] J. Simone et al. [Fermilab Lattice and MILC Collaborations]. „The decay constants f_{D_s}, f_{D^+}, f_{B_s} and f_B from lattice QCD.“ *PoS* **LAT2010** (2010), p. 317.

[Ski06] J. Skilling. „Nested sampling for general Bayesian computation.“ *Bayesian Analysis* **1**.4 (2006), pp. 833–860.

[SS05] D. Scott and S. Sain. „Multidimensional density estimation.“ *Handbook of Statistics* **24** (2005), pp. 229–261.

[SS06] D. S. Sivia and J. Skilling. *Data analysis: a Bayesian tutorial*. Oxford University Press, 2006.

[Tev09] Tevatron Electroweak Working Group. „Combination of CDF and D0 Results on the Mass of the Top Quark“ (2009). arXiv:`0903.2503`.

[Tie94] L. Tierney. „Markov chains for exploring posterior distributions.“ *Ann. Stat.* **22**.4 (1994), pp. 1701–1728.

[Ush+06] Y. Ushiroda et al. [Belle Collaboration]. „Time-Dependent CP Asymmetries in $B^0 \to K^0_S\pi^0\gamma$ transitions.“ *Phys.Rev.D* **74** (2006), p. 111104. arXiv:`hep-ex/0608017`.

[Wal69] A. M. Walker. „On the Asymptotic Behaviour of Posterior Distributions.“ *J.R.Stat.Soc.Series B Stat. Methodol* **31** (1969), pp. 80–88.

[Wei+09] J.-T. Wei et al. [Belle Collaboration]. „Measurement of the Differential Branching Fraction and Forward-Backward Asymmetry for $B \to K^{(*)} l^+ l^-$.“ *Phys.Rev.Lett.* **103** (2009), p. 171801. arXiv:`0904.0770`.

[Win11] M. Wingate. „Lattice QCD Calculations with b Quarks: Status and Prospects.“ *PoS* **BEAUTY2011** (2011), p. 057. arXiv:`1105.4498`.

[Wol83] L. Wolfenstein. „Parametrization of the Kobayashi-Maskawa Matrix.“ *Phys.Rev.Lett.* **51** (1983), p. 1945.

[Wra+09] D. Wraith et al. „Estimation of cosmological parameters using adaptive importance sampling.“ *Phys.Rev.D* **80** (2009), p. 023507. arXiv:`0903.0837`.

[WZ72] K. Wilson and W. Zimmermann. „Operator product expansions and composite field operators in the general framework of quantum field theory.“ *Commun.Math.Phys.* **24** (1972), pp. 87–106.

[Zho+11] R. Zhou et al. [Fermilab Lattice, MILC Collaborations]. „Form Factors for $B \to K\ell^+\ell^-$ Semileptonic Decay from Three-Flavor Lattice QCD.“ *PoS* **LAT2011** (2011), p. 298. arXiv:`1111.0981`.