Inclusive Neutral Current *ep* Cross Sections with HERA II and Two-dimensional Unfolding

Gutachter der Dissertation:	Prof. Dr. Eckhard Elsen
	Prof. Dr. Peter Schleper
Gutachter der Disputation:	Prof. Dr. Jochen Bartels
	Prof. Dr. Eckhard Elsen
Datum der Disputation:	11. Juli 2011
Vorsitzender des Prüfungssausschusses:	Prof. Dr. Caren Hagner
Vorsitzender des Promotionsausschusses:	Prof. Dr. Peter Hausschildt
Dekan der MIN Fakultät:	Prof. Dr. Heinrich Graener
Departmentleiterin des Fachbereichs Physik:	Prof. Dr. Daniela Pfannkuche

Abstract

The electron proton collider HERA in Hamburg is known as a source of accurate data of the proton structure, which is usually described by structure functions, predominantly $F_2\left(x, Q^2\right)$. Since the HERA upgrade in 2001 its luminosity has been significantly increased; at the H1 experiment up to ~ 180 pb^{-1} per year were collected. Today, these data are available for physics analyses, together with the best detector calibration and understanding ever. Such precision data are especially valuable in the light of first LHC data analyses, which need precise knowledge of the proton structure as an important input.

In this thesis, the inclusive neutral current $ep \to eX$ cross section at small e^- scattering angles has been measured using the electromagnetic *SpaCal* calorimeter in the backward region of the H1 detector. This calorimeter constructed of lead and scintillating fiber was designed to measure the scattered electron with high resolution in both energy and polar angle. The analysis comprises the kinematic range of $0.06 < y_e < 0.6$ for the inelasticity and 14 GeV$^2 < Q_e^2 < 110$ GeV^2 for the squared momentum exchange. The data sample consists of positron proton collisions of the years 2006 and 2007, adding up to an integrated luminosity of ~ 141 pb^{-1}.

Due to the high luminosity of the HERA II run phase the accuracy is no longer limited by the data statistics but rather by the detector resolution and systematics. The migration becomes increasingly influential; an effect which leads to distortions of the measured distribution as well as to statistical correlations between adjacent data points. At this stage, the correction of detector effects as well as the precise determination of statistical correlations become important features of a rigorous error treatment.

In this analysis two-dimensional unfolding has been applied. This is a novel approach to H1 inclusive cross section measurements, which are usually based on a bin-by-bin efficiency correction (*bin-by-bin method*). With unfolding, the detector effect to the measurements is modelled by a linear transformation ("response matrix") which is used to correct any distortion of the data. The inclusion of off-diagonal elements results in a coherent assessment of the statistical uncertainties and correlations. The model dependence can be optimally evaluated. In this context, the bin-by-bin method can be viewed as an approximation based on a diagonal response matrix.

In a scenario of limited detector resolution, the unfolded data distributions will typically exhibit strong fluctuations and correlations between the data points. This issue can be adressed by smoothing procedures (*regularization*). Different methods have been tested in the analysis. Among those are an algebraic method, a method exploiting the notion of global correlation and the standard L curve method. All three methods give similar results, which are consistent with the result from the standard bin-by-bin efficiency correction. However, the statistical uncertainties from unfolding are larger than those from the standard bin-by-bin method for all tested regularization prescriptions. For the algebraic method, the statistical uncertainty is of the order of $1 - 2\%$ and the total error of the order of $2 - 3\%$ throughout the kinematic range of this analysis.

The statistical uncertainties from unfolding and from the bin-by-bin method has been compared. This is done by choosing a rather strong smoothing prescription for the unfolding, which leads to a minimum of correlations between the data points. A difference of the order of $20 - 30\%$ in the error is found. This reflects the additional effect of migration on the statistical error, a contribution that has not been accounted for by the bin-by-bin method.

To summarize, the propagation from the standard bin-by-bin efficiency correction to a full two dimensional unfolding treatment does not result in an observable change of the measured cross sections, thus establishing trust in previous measurements. However, a significant impact on the statistical uncertainties is observed, which seem to be clearly underestimated by the traditional (bin-by-bin) error treatment.

Zusammenfassung

Der Elektron-Proton-Beschleuniger HERA in Hamburg ist als Quelle von präzisen Daten über die Protonstruktur bekannt. Diese wird üblicherweise mittels Strukturfunktionen beschrieben, allen voran $F_2\left(x, Q^2\right)$. Seit dem HERA Upgrade im Jahre 2001 wurde die Luminosität signifikant gesteigert; beim H1 Experiment wurden bis zu 180 pb^{-1} pro Jahr erreicht. Diese Daten sind heute für Analysen verfügbar; gleichzeitig sind durch sie maßgebliche Fortschritte beim Detektorverständnis und der Kalibrierung erzielt worden. Diese Präzisionsdaten sind insbesondere deswegen interessant, da die Protonstrukturfunktionen einen wichtigen Input für erste Datenanalysen am LHC darstellen.

In dieser Doktorarbeit wird der inklusive Wirkungsquerschnitt für Elektron-Proton-Streuung $ep \rightarrow eX$ für neutrale Ströme bei kleinen e^--Streuwinkeln gemessen. Dabei wird das elektromagnetische Kalorimeter *SpaCal* im hinteren Bereich des H1 Detektors genutzt. Dieses Blei/Glasfaser-Kalorimeter ist speziell für den Nachweis des gestreuten Elektrons ausgelegt und verfügt über eine hohe Energie- und Polarwinkelauflösung. Die Analyse umfaßt einen kinematischen Bereich von $0.06 < y_e < 0.6$ für die Inelastizität und $14 \text{ GeV}^2 < Q_e^2 < 110 \text{ GeV}^2$ für den quadrierten Impulsübertrag. Dabei wurden Proton-Positron-Kollisionen aus den Jahren 2006 und 2007 wurden untersucht, insgesamt entspricht das einer Luminosität von ~ 141 pb^{-1}.

Aufgrund der hohen Luminosität der HERA II Run Phase ist die Meßgenauigkeit nicht mehr durch die Datenstatistik limitiert, sondern vor allem durch die Auflösungseffekte und systematische Unsicherheiten. Die Migration wird daher immer bedeutsamer; ein Effekt, der zur Verzerrung der gemessenen Verteilung sowie zu statistischen Korrelationen zwischen benachbarten Datenpunkten führt. Auf dieser Ebene wird die Korrektur von Detektoreffekten und die präzise Bestimmung der statistischen Korrelationen immer mehr zu einem wichtigen Bestandteil einer sauberen Fehlerbestimmung.

In dieser Analyse wird zweidimensionales Entfalten angewendet. Für inklusive H1 Wirkungsquerschnitte ist dieses Verfahren neu; normalerweise wird eine bin-weise Effizienzkorrektur verwendet (*Bin-by-Bin-Methode*). Beim Entfalten werden der Effekt des Detektors auf die Messung durch eine lineare Transformation ("Antwortmatrix") modelliert, mit deren Hilfe alle Verzerrungen korrigiert werden können. Die Berücksichtigung der nicht-diagonalen Matrixelemente ermöglicht die kohärente Bestimmung der statistischen Unsicherheiten und Korrelationen. Die Modellabhängigkeiten können optimal bestimmt werden. Die Bin-by-Bin-Methode kann in diesem Zusammenhang als Näherung betrachtet werden, die auf einer diagonalen Antwortmatrix basiert.

Im Fall begrenzter Detektorauflösung zeigt das Entfaltungsergebnis typischerweise starke Fluktuationen und Korrelationen zwischen den Datenpunkten. Diesem Problem kann mit Glättungsverfahren begegnet werden (Regularisierung). Verschiedene Verfahren werden in der Analyse getestet. Darunter sind eine algebraische Methode, eine Methode, die auf dem Begriff der globalen Korrelation beruht, und die L-Kurven-Methode. Alle drei Methoden führen zu ähnlichen Ergebnissen, die auch konsistent mit dem Ergebnis der Bin-By-Bin-Methode sind. Allerdings sind die statistischen Unsicherheiten nach Entfaltung grösser als die von der Bin-By-Bin-Methode und zwar unabhängig von der Regularisationsvorschrift. Für die algebraische Methode ist der statistische Fehler typischerweise $1 - 2\%$ und der Gesamtfehler etwa $2 - 3\%$ im kinematischen Bereich der Analyse.

Die statistischen Fehler von der Entfaltungsmethode und der Bin-By-Bin-Methode werden verglichen. Dafür wird eine starke Glättungsvorschrift für die Entfaltung gewählt, welche die Korrelationen zwischen den Datenpunkten minimiert. Es zeigt sich eine Differenz von etwa $20 - 30\%$ im Fehler. Dieser zusätzliche Effekt beruht auf der Migration, die bei der Bestimmung des Fehlers nach der traditionellen Bin-By-Bin-Methode nicht berücksichtigt wird.

Zusammenfassend gilt, dass der Übergang von der klassischen Bin-By-Bin-Methode zur vollen zweidimensionalen Entfaltungsmethode zu keinen beobachtbaren Änderungen an den Wirkungsquerschnitten führt. Dagegen ist der Einfluss auf die statistischen Unsicherheiten signifikant. Diese werden durch die klassische Bin-By-Bin-Methode unterschätzt.

Contents

Chapter 1

Introduction and Motivation

1.1 DIS and Hadron Structure

One of the essential experimental tools to determine the internal structure of hadrons such as the proton is known as *Deep Inelastic Scattering* (DIS). Experiments of this type probe the hadrons with fundamental leptons such as electrons, myons or neutrinos, which interact only via the electroweak force. The scattering process can thus be regarded independently from the proton structure, which itself results from the strong force. An introduction to DIS can be found in [ES] and [CS03].

In this picture the DIS cross section can be expressed in terms of structure functions, which contain all information of the hadron structure to which the process at hand provides sensitivity. For neutral current processes, i. e. processes that preserve the charge of the probing particle, the hadron structure is described by the functions $F_2(x, Q^2)$ and $F_L(x, Q^2)$. Here, the variables x and Q^2 describe the kinematics of the scattering process, see section 2.1. Different structure functions are obtained for charged current processes.

The theory used to describe the structure of hadrons is *Quantum Chromo Dynamics* (QCD). This highly successful $SU(3)$-invariant gauge theory comprises of massive fermions called *quarks* and massless exchange bosons called *gluons*. The quarks come in six flavours, called *up* (u), *down* (d), *charm* (c), *strange* (s), *top* (t) and *bottom* (b) and occur as particle or antiparticle. Both quarks and gluons carry *color charge*, which can take on the values *red*, *green* and *blue*. The fact, that gluons carry color charge gives rise to the phenomenon of strong self-interaction. For a general introduction to Quantum Chromo Dynamics see [MEP95] and [Wei05b].

In the QCD picture, each hadron is regarded as a bound state of quarks and gluons, which in this context are also called *partons*. To describe the hadron structure in terms of these QCD objects one introduces a *Parton Density Function* (PDF) $f_i(x, Q^2)$ for every type i of parton, be it a quark, antiquark or gluon [ES]. The parton density functions provide a picture of the proton that is independent of the specifics of a scattering process. Of course, these functions are related to the structure functions. In fact, the PDFs can be extracted by a simultanous fit to a given set of measured structure functions (*QCD Fit*) [A$^+$10].
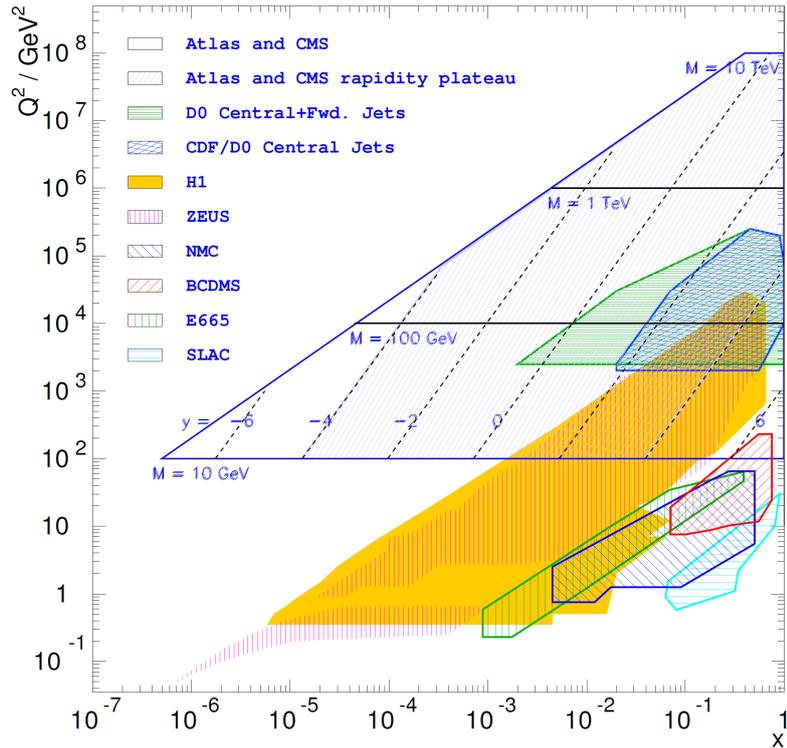
Figure 1.2.1: *Kinematic range of experiments dedicated to proton structure determination. Note the large coverage of the HERA experiments H1 and ZEUS, ranging up to $Q^2 \simeq 2 \cdot 10^4$ GeV^2 and down to $x \simeq 0.5 \cdot 10^{-6}$. The kinematic range of future measurements at the LHC is also indicated.*

1.2   HERA and the *ep*-Experiments

One of the best studied hadrons is certainly the proton. Ever since the discovery of scaling for *ep*-scattering at the Stanford Linear Accelerator Center in 1969 (see chapter 2), numerous μ- and ν-scattering experiments were performed to determine the proton structure. Among them are the fixed target experiments NMC and BCDMS [A+97b, B+89].

The HERA[1] accelerator in Hamburg was the first *ep* collider ever built and significantly increased our knowledge about the structure functions $F_2(x, Q^2)$ and $F_L(x, Q^2)$. While fixed target experiments were able to resolve the proton structure only up to $Q^2 \simeq 2 \cdot 10^2$ GeV^2 and $x \simeq 10^{-2}$, at HERA this range was extended to $Q^2 \simeq 2 \cdot 10^4$ GeV^2 and $x \simeq 0.5 \cdot 10^{-6}$, see figure 1.2.1. This was achieved at the two 4π detectors H1 [A+97a] and ZEUS [Col93], at which collisions at a center of mass energy of up to 318 GeV were studied. Figure 1.2.2 summarizes data from NMC, BCDMS and the two HERA experiments.

From 1992 to 2007, both HERA and ZEUS collected a total integrated luminosity of nearly 500 pb^{-1}. The majority of this data sample originates from the HERA II run phase (2001 - 2007) and particularly from its last two years of operation, in which the accelerator reached its optimal performance of nearly $\sim$ 180 pb^{-1} per year, see also figure 1.2.3. Today, these data are available for physics analyses, together with the best

---

[1]Hadron-Elektron-Ring-Anlage
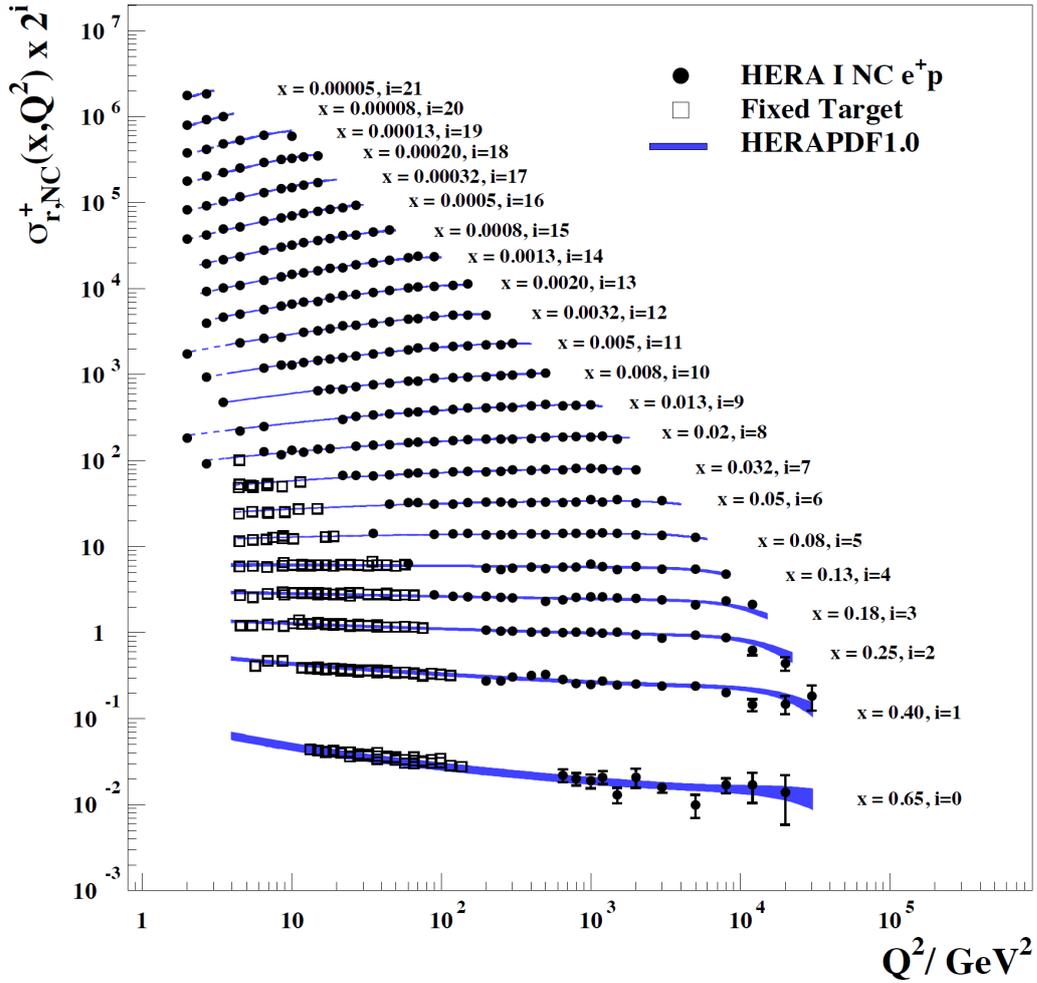
Figure 1.2.2: *Proton structure measured at HERA. The reduced neutral current* $ep \rightarrow eX$ *cross section measured at H1 and ZEUS is shown, together with measurements from the fixed target experiments NMC and BCDMS [A^+97b, B^+89]. The blue band corresponds to NLO DGLAP calculations based on HERAPDF1.0 [A^+10]. (Note, that the HERA data in this plot are obtained with the bin-by-bin method.)*
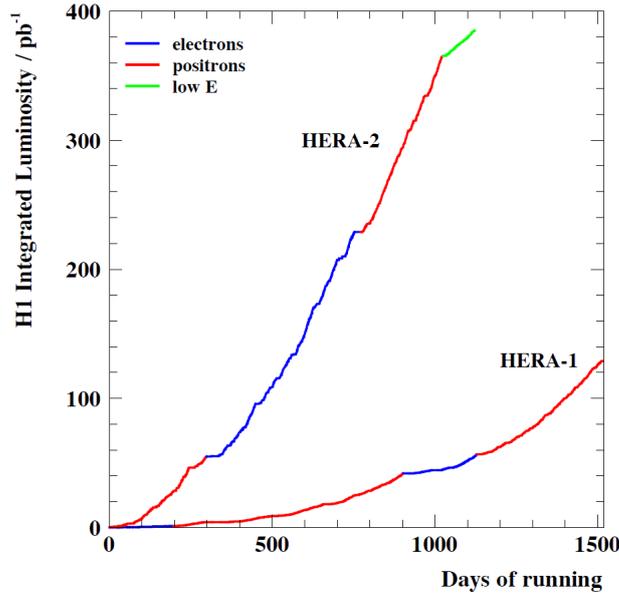
Figure 1.2.3: *Total integrated luminosity $\mathcal{L}_{int}$ measured at the H1 detector as a function of time. The best performance was reached in the last years of operation (HERA II). The total integrated luminosity reaches nearly ~ 500 pb^{-1}.*

detector calibration and understanding ever. In fact, they challenge the experiments to measure the F_2 proton structure function with an experimental uncertainty of $\sim 1\%$ throughout most of the HERA phase space.

In this thesis, the inclusive neutral current $ep \rightarrow eX$ cross section has been measured using the electromagnetic SpaCal calorimeter. This lead-based calorimeter was designed to detect the scattered electron at small scattering angles with high energetic and spatial resolution. The analysis comprises the kinematic range of $0.06 < y_e < 0.6$ for the inelasticity and 14 GeV$^2 < Q_e^2 < 110$ GeV^2 for the squared momentum exchange. The data sample consists of positron proton collisions from 2006 and 2007, adding up to an integrated luminosity of ~ 141 pb^{-1}.

1.3   The Importance of HERA for the LHC

The precise understanding of the proton structure is particularly relevant for analyses at the Large Hadron Collider (LHC) [CS10, Pet10]. The LHC is a proton proton collider and consequently every cross section prediction at the LHC depends inherently on the proton structure. In fact, the dependance is even "quadratic", if the proton structure is expresses in terms of parton density functions.

From figure 1.2.1 it can be seen, that the kinematic range in the direction of the Bjorken variable x is nearly the same for HERA and LHC. Therefore, the HERA data have a significant influence on every LHC analysis. This constitutes the need for utmost accuracy in F_2 measurements.

1.4 Limitations of the Bin-By-Bin-Method

Best accuracy, however, is only achieved on the basis of a thorough and concise understanding of detector effects. Every measurement in experimental physics suffers from the limited resolution and the limited efficiency of the measuring device. During the measurement process, events can be reconstructed in the wrong bin (*migration effects*) or can be lost completely (*inefficiency*). This is particularly true for steeply varying distributions. Therefore, the measured distribution may be distorted and statistical correlations between data points may arise.

Traditionally, inclusive cross section measurements at H1 have used a mere bin-by-bin efficiency correction (*bin-by-bin method*). This method essentially compares the generated and "measured" event counts for a Monte Carlo simulation, thus establishing a bin-wise correction factor. If the Monte Carlo simulation can be trusted, this method sufficiently corrects for efficiency effects. In addition, it corrects distortions in the shape of the measured distribution which arise from migration effects.

However, the bin-by-bin method cannot account for statistical correlations that arise between adjacent data points due to migration effects. Effectively, these correlations are ignored. A QCD fit may be altered by such an artificial prescription. Moreover, the statistical uncertainties on the data points are expected to rise in the presence of migration effects. Again, this effect is not accounted for by the bin-by-bin method.

Note, that in a scenario of low event statistics, the bin-by-bin method may be sufficient. Migration effects strongly depend on the choice of the bin grid. Imagine a binning that is coarse in comparison to the detector resolution, as is a typical case if the choice of the bin grid is driven by low statistics. Then, the migration effects will be rather small, potentially even negligible and the bin-by-bin method can be used safely. However, with increasing event statistics the choice of the binning will eventually be driven by the detector resolution, which becomes the limiting factor in the experimental accuracy. Therefore, migration effects will play a larger role with high statistics and the bin-by-bin method will no longer be applicable.

The HERA II data sample provides an amount of data, which indeed makes the detector resolution the limiting factor for the bin grid. Maximizing the information output from the measurement means choosing a bin grid which reflects the local detector resolution. (We will discuss these issues at length in chapter 4.) A full treatment of migration effects becomes necessary.

1.5 Unfolding

A concise way to address both efficiency and migration effects is unfolding. This statistical tool models detector effects by a linear transformation ("response matrix") which in turn is used to correct distortions of the data. Moreover, detector-based correlations between the data points are quantified. In chapter 4 we review some technical aspects of unfolding.

A typical feature of unfolding is the appearance of large fluctuations between the bins. Such correlations can be reduced by the introduction of a smoothing procedure, called *regularization*. At the price of potentially introducing a bias, the measurement is thus freed from unphysical correlations. Evidently, a reasonable trade-off needs to

be found. The determination of the adequate level of regularization is a key challenge of each unfolding procedure.

1.6   How Unfolding effects H1 Cross Sections

In this analysis, two dimensional unfolding in y and Q^2 is employed. This is a novel approach to inclusive cross section measurements at H1, which are usually applied the mere bin-by-bin efficiency correction discussed above. The application of unfolding results in a concise assessment of the statistical uncertainties, as well as a first-time natural evaluation of bin-to-bin correlations. Moreover, a reduction of model dependencies can be achieved, leading to a better understanding and reduction of systematic uncertainties.

Of course, strong interest is directed towards the potential variations between the results obtained from unfolding and from the bin-by-bin method. We will see in chapter 10, that no discrepancy in the shapes of the measured distributions is observed. However, effects in the statistical uncertainties do emerge and depend on the level of regularization.

In the light of all this, this work can also be seen as cross check of the standard H1 error treatment. Note, that for a conclusive comparison of statistical uncertainties uncorrelated quantities have to be studied. This condition can be approximately fulfilled by choosing a level of regularization that minimizes the global correlation, see chapter 4. A difference of the order of $20 - 30\%$ is found for the statistical uncertainties. This reflects the effect of migration on the statistical error, a contribution that has not been accounted for by the bin-by-bin method.

Chapter 2

DIS and QCD

In this chapter some basic concepts of Deep Inelastic Scattering (DIS) and the proton structure are reviewed. We introduce the HERA event kinematics, the parton model of the proton and the QCD DGLAP evolution. Moreover, QED radiative effects are discussed. See also [ES] for a general introduction into DIS and QCD.

2.1 Event kinematics

In figure 2.1.1, a typical *neutral current interaction* $ep \rightarrow eX$ is depicted. Let the symbols $\boldsymbol{k}$ and $\boldsymbol{k}'$ denote the four momenta of the incoming and outgoing lepton, $\boldsymbol{P}$ the four momentum of the incoming proton. At fixed center of mass energy, the event kinematics is typically described by the following Lorentz invariant variables:

$$Q^2 \equiv -q^2 := -\left(\boldsymbol{k} - \boldsymbol{k}'\right)^2 \tag{2.1.1}$$

$$x := \frac{Q^2}{2\boldsymbol{Pq}} \tag{2.1.2}$$

$$y := \frac{\boldsymbol{qP}}{\boldsymbol{kP}} \tag{2.1.3}$$

Here, Q^2 is known as *squared momentum exchange*, y as *inelasticity* and x as *Bjorken variable.* Both x and y are by construction limited to values between 0 and 1. In the quark parton model, x corresponds to the longitudinal momentum fraction carried by the struck parton. Two more important Lorentz invariants are the center of mass energy $\sqrt{s}$ of the ep-system and the center of mass energy W of the proton-boson system:

$$s := \left(\boldsymbol{k} + \boldsymbol{P}\right)^2 \tag{2.1.4}$$

$$W^2 := \left(\boldsymbol{q} + \boldsymbol{P}\right)^2 \tag{2.1.5}$$

Since the masses m_P and m_e of proton and electron are negligible compared to the HERA energy scale $\sqrt{s}$, we directly obtain the relations:

$$s = \sqrt{4E_p E_e} \tag{2.1.6}$$

$$Q^2 = sxy \tag{2.1.7}$$

$$sy = W^2 + Q^2 \tag{2.1.8}$$

For more information on HERA event kinematics, see [Sti01].
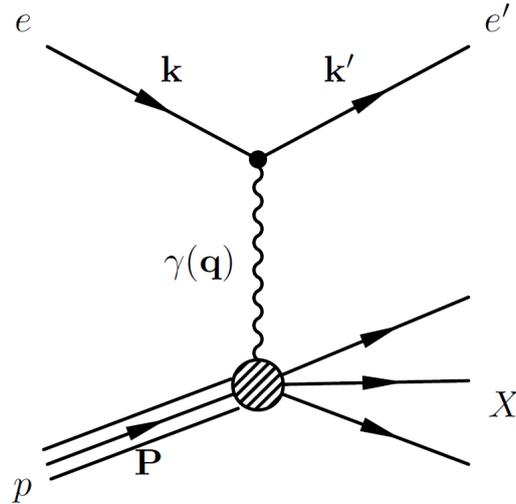
Figure 2.1.1: *Neutral current $ep \rightarrow eX$ DIS event in lowest order. The symbols $\boldsymbol{k}$, $\boldsymbol{k}'$ and $\boldsymbol{P}$ denote the four-momenta of the incoming and outgoing electron and the incoming proton, respectively. The gauge boson is a photon with four momentum $\boldsymbol{q}$; for high Q^2 contributions from the Z_0 boson and interference terms have to be accounted for.*

2.2   DIS Cross Section and Structure Functions

Due to momentum conservation, ϕ-symmetry and negligible particle masses ($m \simeq 0$), the double differential $ep \rightarrow eX$ cross section can only depend on three variables. Without loss of generality, these are s, x and y. However, during an experimental run phase at HERA, the center of mass energy $\sqrt{s}$ is fixed.

For $Q^2 \lll m_Z^2$, the scattering process is mediated purely by γ-exchange, thus the cross section can be calculated from *Quantum Electro Dynamics* (QED). By exploitation of Lorentz invariance and electromagnetic current conservation we obtain the following general structure [ES]:

$$\frac{\partial^2}{\partial x \partial Q^2} \sigma^{NC}\left(x, Q^2\right) = \frac{2\pi\alpha^2 Y_+(y)}{xQ^4} \left(F_2\left(x, Q^2\right) - \frac{y^2}{Y_+(y)} F_L\left(x, Q^2\right) \right) \qquad (2.2.1)$$

Here, $F_2\left(x, Q^2\right)$ and $F_L\left(x, Q^2\right)$ are the *proton structure functions*. They encode the internal substructure of the proton, which cannot be derived from perturbative QCD and therefore has to be determined experimentally. Note the abbreviation $Y_+(y) := 1 + (1-y)^2$. Moreover, α denotes the electromagnetic fine structure constant. Note, that for $Q^2 \gtrsim m_Z^2$, terms for the Z-exchange and for γ/Z-interference need to be added to (2.2.1).

The function $F_L\left(x, Q^2\right)$ is called *longitudinal structure function*. For low y, its contribution to the cross section is suppressed with $\frac{y^2}{Y_+(y)}$. For $y \lesssim 0.6$, the measurement of the neutral current cross section can therefore be regarded as a direct F_2 measurement up to a correction of the order of a few percent.

For brevity, one typically quotes the *reduced cross section* σ_r by suppressing the

kinematic factor in (2.2.1):

$$\sigma_r\left(x,Q^2\right) := F_2\left(x,Q^2\right) - \frac{y^2}{Y_+(y)} F_L\left(x,Q^2\right) \tag{2.2.2}$$

2.3   The Quark Parton Model

In the *infinite momentum frame*, the proton is viewed as being composed of non interacting constituents called *partons*. Each parton carries a fraction $\xi < 1$ of the proton's longitudinal momentum, while its transverse momenta are neglected. The cross section (2.2.1) is then given as an incoherent sum over the interaction processes with each individual parton. This gives:

$$\frac{\partial^2}{\partial x \partial Q^2}\sigma^{NC}\left(x,Q^2\right) = \sum_i \int_0^1 d\xi f_i\left(\xi\right) \times \frac{\partial^2}{\partial x \partial Q^2}\sigma^{NC}\left(x,Q^2\right)\bigg|_{eq_i \to eq_i} \tag{2.3.1}$$

Here, $f_i\left(\xi\right)$ is the *Parton Density Function* of the parton of type i.

In a very simple model, the proton consists of three non-interacting fermions, called *valence quarks*, namely two u- and one d-quark. For this scenario, we directly obtain from QED:

$$\frac{\partial^2}{\partial x \partial Q^2}\sigma^{NC}\left(x,Q^2\right)\bigg|_{eq_i \to eq_i} = \frac{2\pi\alpha^2 Y_+(y)}{Q^4} \cdot e_i \delta(x - \xi) \tag{2.3.2}$$

Here, e_i denotes the charge of the i-th quark. By insertion of (2.3.2) in (2.3.1) we directly obtain $x = \xi$ and moreover

$$F_2\left(x,Q^2\right) = \sum_i e_i^2 x f_i\left(x\right) \tag{2.3.3}$$

$$F_L\left(x,Q^2\right) = 0 \tag{2.3.4}$$

Note the vanishing Q^2-dependence in (2.3.3), a feature called *scaling* [Bjo69]. The vanishing longitudinal structure function (*Callan Gross relation*) reflects the fermionic nature of the quarks. Both features were predicted in 1969 [Bjo69, CG69] and were measured in 1972 [M+72]. Historically, they played a crucial role in establishing the quark picture in DIS.

2.4   DGLAP Evolution

Of course, a proton model of non-interacting partons can only be an approximation. In fact, violations of the scaling law have been found rather soon [F+74]. A more rigorous treatment is needed, which takes into account the QCD interactions. These are the emission of gluons from quarks and from other gluons, the splitting of gluons in quark-antiquark pairs and the gluon self-interaction. The diagrams are shown in figure 2.4.1.

For a quantiative QCD treatment of the neutral current *ep* cross section, we start from the *factorization theorem*, see [CSS88]. It states, that a cross section is composed

of a *short distance part* describing the hard scattering process and a *long distance part* describing the proton structure. For F_2, we can write:

$$F_2\left(x, Q^2\right) = \sum_i \int_z^1 dz \, C_2^i \left(\frac{x}{z}, \frac{Q^2}{\mu_r^2}, \frac{\mu_f^2}{\mu_r^2}\right) f_i\left(z, \mu_r^2, \mu_f^2\right) \tag{2.4.1}$$

Here, the sum runs over all parton types: quarks, antiquarks and gluons. Again, f_i denote parton density functions and C_2^i describe the hard scattering process. With μ_r and μ_f we denote the *renormalization scale* and *factorization scale.*

Note, that F_2 must ultimately not depend on the choice of the scales μ_f and μ_r. In the *DIS renormalization scheme* we conveniently choose $\mu_r = \mu_f \equiv \mu$. Then, (2.4.1) simplifies to:

$$F_2\left(x, Q^2\right) = \sum_i e_i^2 x \cdot f_i\left(x, Q^2\right) \tag{2.4.2}$$

Moreover, by exploiting the arbitrariness of the renormalization scheme μ, one can derive the DGLAP equations [AP77, Dok77, GL72a, GL72b]:

$$\frac{\partial q_i\left(x, Q^2\right)}{\partial \ln Q^2} = \frac{\alpha_s\left(Q^2\right)}{2\pi} \int_x^1 \frac{dz}{z} \left[\sum_j q_j\left(z, Q^2\right) P_{ij}\left(\frac{x}{z}\right) + g\left(z, Q^2\right) P_{ig}\left(\frac{x}{z}\right)\right] \tag{2.4.3}$$

$$\frac{\partial g\left(x, Q^2\right)}{\partial \ln Q^2} = \frac{\alpha_s\left(Q^2\right)}{2\pi} \int_x^1 \frac{dz}{z} \left[\sum_j q_j\left(z, Q^2\right) P_{gj}\left(\frac{x}{z}\right) + g\left(z, Q^2\right) P_{gg}\left(\frac{x}{z}\right)\right] \tag{2.4.4}$$

Here, $q_i\left(x, Q^2\right)$ denote the quark and antiquark density functions, $g\left(x, Q^2\right)$ the gluon density function and $\alpha_s\left(Q^2\right)$ the running coupling of the strong force. The symbols $P_{ij}(z)$, $P_{gj}(z)$, $P_{ig}(z)$ and $P_{gg}(z)$ are called *splitting functions* which can be calculated perturbatively in $\frac{\alpha_S}{2\pi}$. Calculations up to NNLO are available [VVM05, MVV04, VMV04]. Their first order terms $P_{ba}^{(0)}(x/z)$ measure the probability for a parton of type a with momentum fraction z to emit a parton of type b with momentum fraction x, see figure 2.4.1.

Note, that equations (2.4.3) and (2.4.4) establish a Q^2-dependence of the parton density functions which explains the observed scaling violations. While for a given Q_0^2 the curve progression of $f_i\left(x, Q^2\right)$ with x can only be measured experimentally, the DGLAP equations allow a theoretical extrapolation to higher Q^2. Typically, a starting value of $Q_0^2 \simeq 2 - 4$ GeV^2 is chosen.

In figure 1.2.2 an overview of F_2-measurements from H1 and ZEUS is given, together with data points from fixed target experiments. The blue band corresponds to a QCD fit, which is referred to as HERAPDF1.0 [A$^+$10]. Note the approximate scaling behavior for $x \sim 0.2$.

2.5   Radiative Effects

A neutral current cross section measurement will be distorted by higher order QED processes, henceforth regarded as *radiative effects* [Spi92]. The most prominent contributions originate from photon radiation off the electron line, see figure 2.4.2. The propagator structure of these graphs suggests, that significant contributions only occur, if at least one of the photon virtualities q^2, q'^2 or q''^2 vanishes. This allows the following classification:
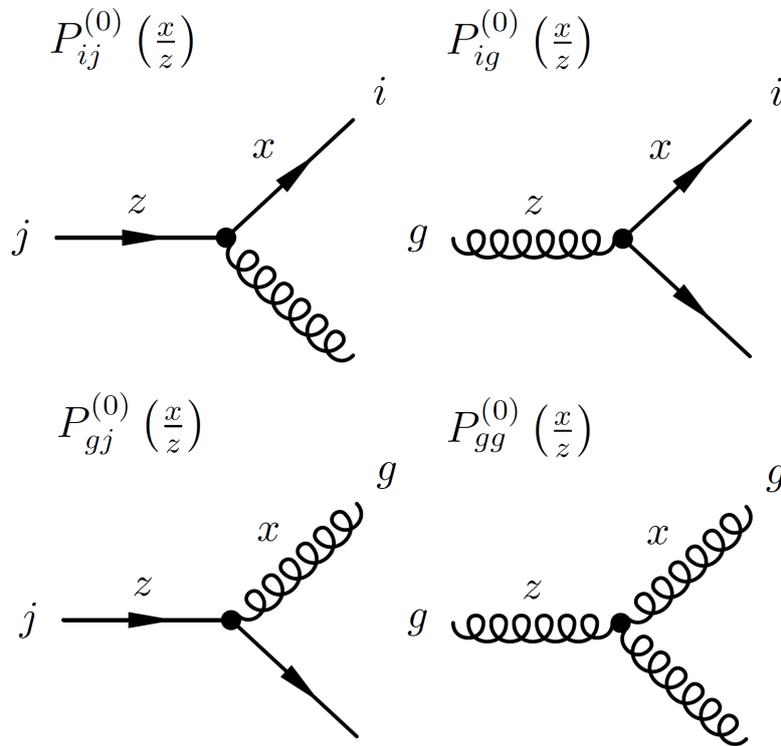
Figure 2.4.1: *Lowest order diagrams for the* splitting functions *in the DGLAP evolu-tion. The first order terms of the splitting functions $P_{ba}^{(0)}(z)$ measure the probability for a parton of type a with momentum fraction z to emit a parton of type b with momentum fraction x.*
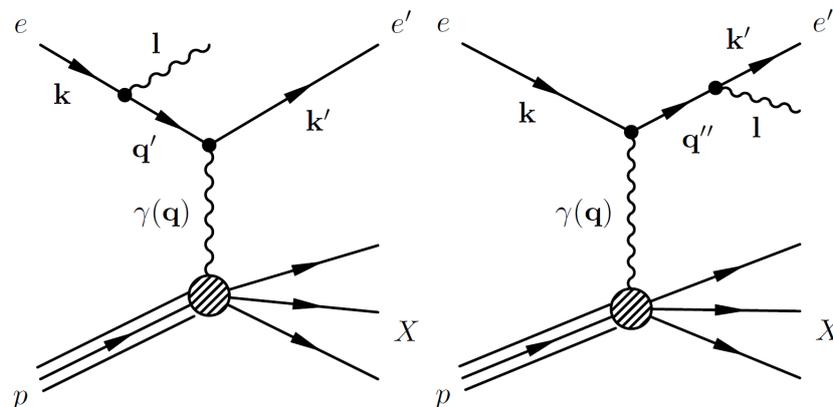
Figure 2.4.2: *Diagrams for Initial State Radiation (ISR) and Final State Radiation (FSR).*

Initial state radiation (ISR). This process is characterized by a single electron being emitted collinearly from the incoming electron with $q'^2 \simeq 0$ and $q^2 \neq 0$. While the photon is typically not detected, the center of mass energy s at the hadronic vertex is effectively lowered which in turn effects the cross section.

Final state radiation (FSR). This process is characterized by a single electron being emitted collinearly from the outgoing electron with $q''^2 \simeq 0$ and $q^2 \neq 0$. This process is less critical, since the emitted photon can typically not be resolved from the electron in the calorimeter and both particles contribute to the cluster energy.

QED Compton (QEDC). This process is characterized by $q^2 \simeq 0$ and $q'^2 \neq 0$ or $q''^2 \neq 0$, respectively. While the proton is typically lost, electron and photon have sizable transverse momentum. A clear signature of two clusters back to back in azimuth and little hadronic activity make these events easy to identify.

Bethe Heitler Events. This process is characterized by $q^2 \simeq 0$ and $q'^2 \simeq 0$ or $q''^2 \simeq 0$, respectively. Since the elastic QED cross section of this process is known to high precision, these events are used for the luminosity determination, see chapter 3.

In a scattering experiment photon radiation can be a substantial effect. However, our interest is mainly directed to the *Born cross section*, which describes the *ep* scattering without radiative effects. Therefore, a correction has to be applied to the measurements, an issue we will re-adress in subsection 5.1.4 and chapter 11.

Chapter 3

The Detector

The objective of this chapter is the introduction of the HERA collider and the H1 detector. We discuss the subdetectors that are relevant to this analysis. Moreover, a brief introduction to the trigger system is given. For a more general description of the H1 detector, see [A+97a].

3.1 HERA Collider

The HERA collider was an electron proton collider on the DESY site in Hamburg, which was operational from 1992 to 2007. It was situated in a circular tunnel of 6.3 km circumference. In figure 3.1.1 the accelerator with its four experiments is shown. H1 and ZEUS were designed at 4π-detectors, at which the electron and proton beam were brought into collision at zero angle. HERMES and HERA-B were fixed target experiments.

In the year 2000 the collider has been shut down for a luminosity upgrade. This divides the HERA run time into two operational phases, HERA I and HERA II. Moreover, different *run periods* exist, which differ in the electron charge (e^-/e^+). For most HERA II run periods, the proton and electron beam energies were $E_{beam}^p = 919$ GeV and $E_{beam}^e = 27.6$ GeV, respectively. Accordingly, the center of mass energy was:

$$s = \sqrt{4E_{beam}^p E_{beam}^e} = 318.5 \text{ GeV} \qquad (3.1.1)$$

Each run period is further subdivided into *HERA luminosity fills,* which typically span 12 hours. At the beginning of a fill, the beam currents reached $I_p \simeq 100$ mA and $I_e \simeq 40$ mA.

Both beams are structured into 220 *bunches,* which collide with a frequency of 10.4 MHz or every 96 ns. (Note however, that typically not all of the bunches are filled, and some unpaired *pilot bunches* exist for the purpose of beam monitoring.) The longitudinal shape of the proton bunches is nearly Gaussian with a width of 45 cm (FWHM), leading to a luminous region with a width of $\sigma_z \sim 9$ cm. However, due to technical reasons, small *satellite bunches* occur about 70 cm apart from the main bunch, which are not used for the analysis.
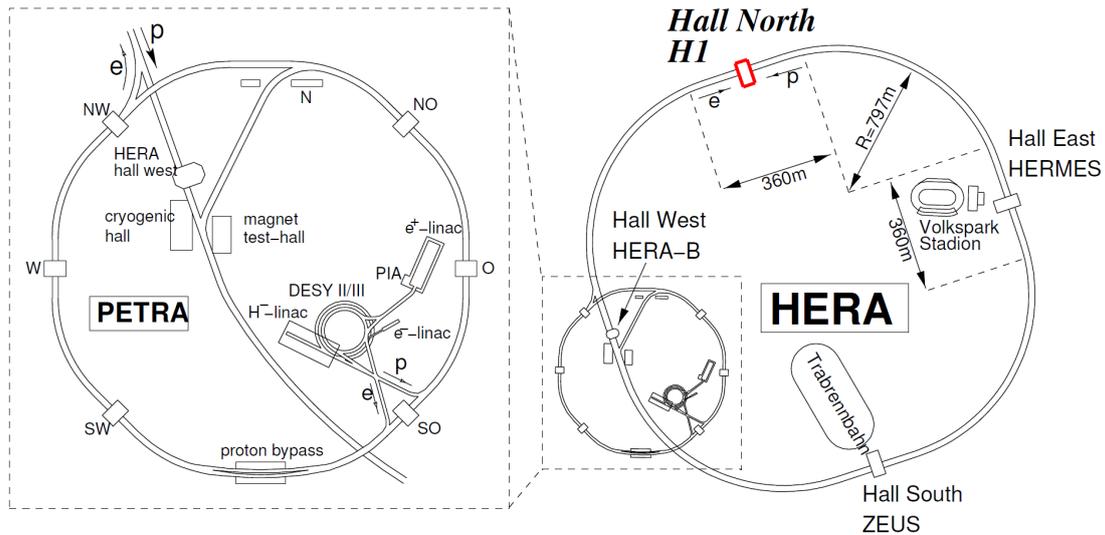
Figure 3.1.1: *HERA with its four experiments H1, ZEUS, HERMES and HERA-B. On the left site, the chain of pre-accelerators is enlarged.*

3.2   The H1 Detector

In figure 3.2.1 an overview of the H1 detector is given. Clearly visible is its asymmetrical structure, which is a consequence of the different beam energies involved. Note the *H1 coordinate system.* It assigns the direction of the incoming proton to the positive z-axis. According to the direction of the z-axis we speak of the *forward region*, the *central region* and the *backward region* of the H1 detector.

The main tracking systems are the *central tracker (CTD)* and the *forward tracker (FTD)*. The most important calorimetry devices are *Liquid Argon Calorimeter (LAr)* and the *SpaCal*, the latter of which is located in the backward H1 region. Both feature electromagnetic and hadronic sections. The calorimetric region is surrounded by a superconducting solenoid which provides a uniform magnetic field of 1.15 T. Outside the solenoid a *myon system* is installed.

3.2.1   The Tracking System

The tracking system consists of the central tracker (CTD) and the forward tracker (FTD). Their total polar angle coverage is $15° \leq \theta \leq 165°$ and $7° \leq \theta \leq 25°$, respectively. Figure 3.2.2 shows an overview of its subcomponents.

Central Jet Chambers CJC1 and CJC2. These concentric drift chambers measure *wire hits* in the $r - \phi-$plane with a spatial resolution of up to 170 μm. This is achieved with 720 and 1920 *sense wires*, respectively, which are oriented parallel to the beam axis. Moreover, a measurement of the z-coordinate is possible ($\delta_z \simeq 4$ cm) by utilization of charge division. The length of the detectors is 220 cm, while the transverse range is 20.3 cm $\leq R \leq$ 45.1 cm for CJC1 and 53 cm $\leq R \leq$ 84.4 cm for CJC2.

Central Outer Z Chamber COZ. This drift chamber consists of radially oriented wires, providing a precise measurement of the z-position of hits with a resolution of $200 - 260\mu m$. It is longitudinally subdivided in 24 rings of 9 cm width. Its radial
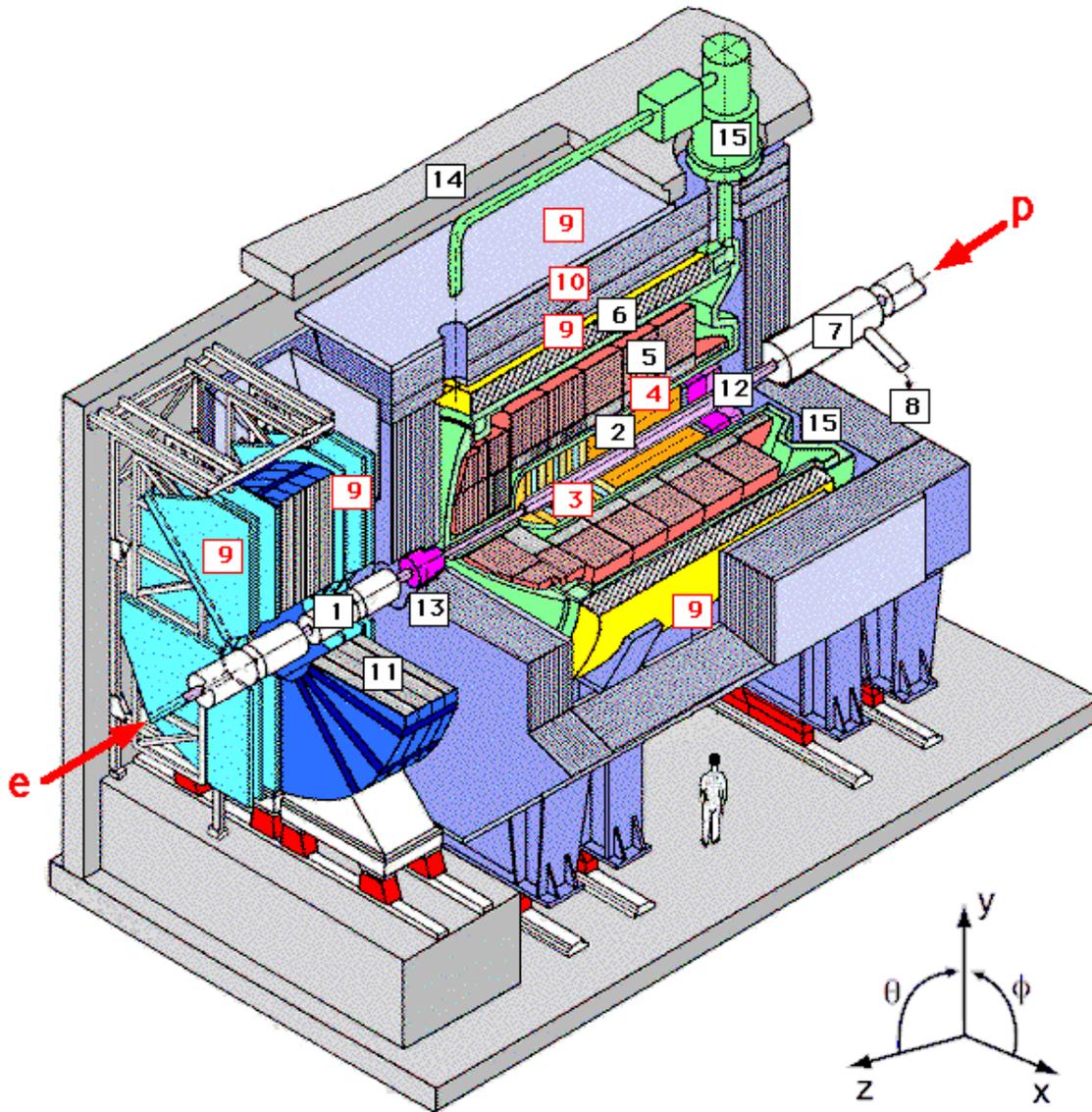
Figure 3.2.1: *The H1 Detector. Important subcomponents: Central Tracker CTD (2), Forward Tracker (3), Electromagnetic LAr (4), Hadronic LAr (5), Superconducting Coil providing 1.15 T (6), Myon System (9), Lead/Scintillating-fiber calorimeter SpaCal (12).*
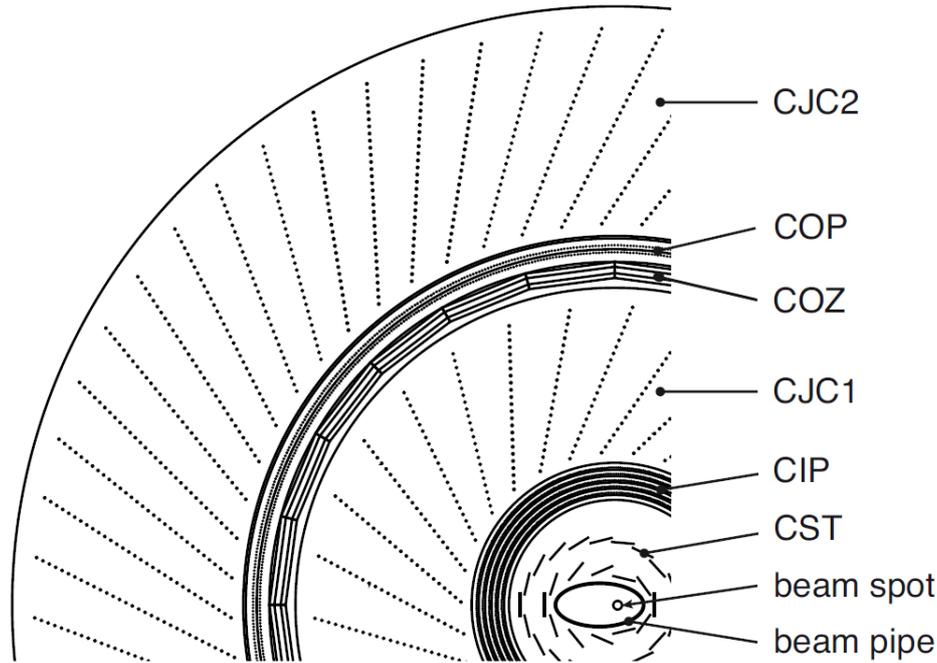
Figure 3.2.2: *Central tracker system (CTD) with its components CST, CIP, CJC1, COZ, COP and CJC2.*

position is between CJC1 and CJC2.

Central Proportional Chambers CIP and COP. Both these chambers are mainly used for triggering and are mounted inside CJC1 and CJC2, respectively. The CIP was redesigned for the HERA II run phase (*CIP2k*), comprising now 5 concentric layers, 8 azimuthal segments and 60 longitudinal sectors of 3.6 cm length each. The spatial hits from the electron track can be used together with its SpaCal cluster to find the longitudinal position z_{Vtx} of the event vertex, see section 3.4.2.

Silicon Track Detectors CST, BST and FST [P+00]. These silicon based strip detectors constitute the innermost layer of the H1 architecture. The CST has a spatial resolution of 12 μm in $r\phi$ and 22 μm in z-direction and supports both the measurement of the event vertex and of potential secondary vertices.

Forward Track Detector (FTD). This tracker system covers the forward detector region. It consists of 3 super modules, each consisting of multiple drift chambers with different wire geometries. Its typical resolution is 210 μm in $r\phi$ and 3 cm in z-direction.

Backward Proportional Chamber (BPC). The BPC is located in front of the SpaCal and consists of six wire layers oriented in 3 different azimuthal directions, see figure 3.2.7. For charged particles, it provides a point measurement in the $xy-$plane with a resolution of 1 mm.

3.2.2   Liquid Argon Calorimeter

The H1 detector features an almost complete 4π calorimetric coverage. While the *SpaCal* covers the backward region ($153° \leq \theta \leq 174°$), the central and forward region is covered by the *Liquid Argon calorimeter (LAr)* ($4° \leq \theta \leq 154°$). Both devices have electromagnetic and hadronic parts.
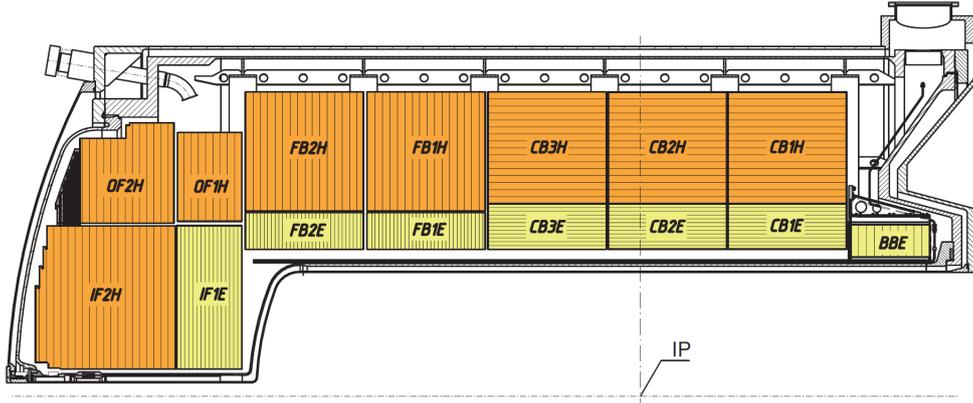
Figure 3.2.3: *Liquid Argon Calorimeter (LAr) with its 8 calorimeter wheels. The inner components (IF1E, FB2E, FB1E, CB3E, CB2E, CB1E, BBE) belong to the electromagnetic part (green), the outer components (IF2H, OF2H, OF1H, FB2H, FB1H, CB3H, CB2H, CB1H) belong to the hadronic part (orange).*

The *Liquid Argon Calorimeter (LAr)* features lead as absorber for the electromagnetic part and stainless steel for the hadronic part. Its cryostat temperature is 90 K. Being placed within the superconducting solenoid, energy loss due to dead material in front of the calorimeter is reduced. The calorimeter is structured in 8 *wheels*, each of which consists of 8 sections in ϕ-direction. Energy resolutions of

$$\frac{\sigma_{E_{EL}}}{E_{EL}} \simeq \frac{10\%}{\sqrt{E_{EL}/\text{GeV}}} \oplus 1\% \tag{3.2.1}$$

and

$$\frac{\sigma_{E_{HAD}}}{E_{HAD}} \simeq \frac{50\%}{\sqrt{E_{HAD}/\text{GeV}}} \oplus 2\% \tag{3.2.2}$$

are reached for test beams [A$^+$93a, A$^+$93b].

3.2.3 The SpaCal calorimeter

The *SpaCal (Spaghetti Calorimeter)* [A$^+$96, N$^+$96] covers the backward region ($153° \leq \theta \leq 174°$). The electromagnetic part is structured into 4.05 cm $\times$ 4.05 cm $\times$ 25 cm *cells,* which are grouped in *super modules* of 4×4 cells each, see figure 3.2.4. Each cell consists of 2340 longitudinally aligned scintillating fibers embedded in a lead matrix, see figure 3.2.6. The fiber diameter is 0.5 mm and the lead/fiber ratio is 2.3 : 1. The fibers conduct the light from showers directly to photo-multiplier tubes (PMT). The stability of the PMT amplification is monitored with pulsed LED signals. A longitudinal cut through the SpaCal is given in figure 3.2.5.

With the electromagnetic SpaCal an extraordinarily accurate energy resolution of

$$\frac{\sigma_{E_{EL}}}{E_{EL}} \simeq \frac{7\%}{\sqrt{E_{EL}/\text{GeV}}} \oplus 1\% \tag{3.2.3}$$
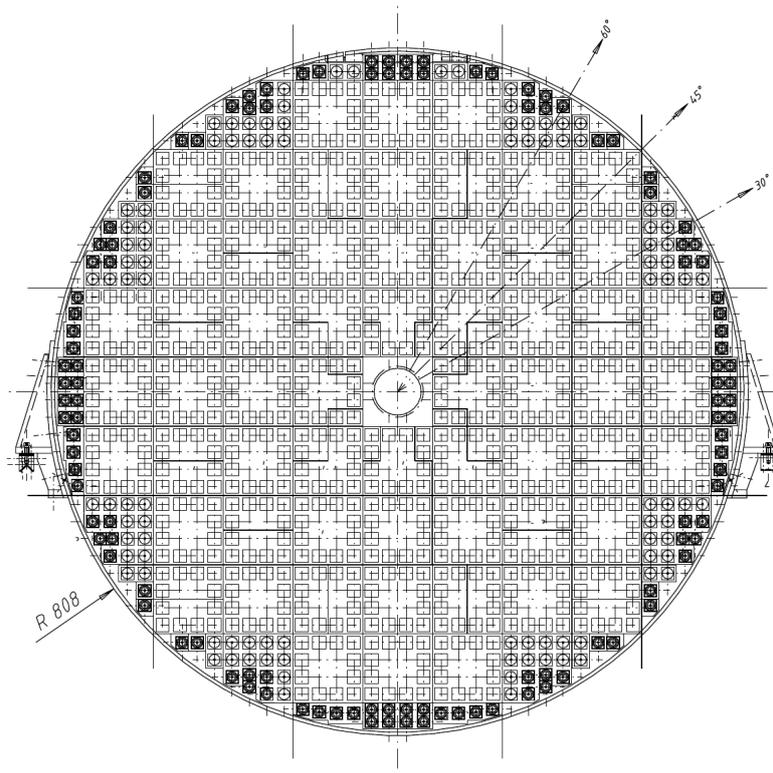
Figure 3.2.4: *The electromagnetic SpaCal calorimeter viewed in transverse projection. Clearly visible are the different* cells *and their grouping in* supermodules.
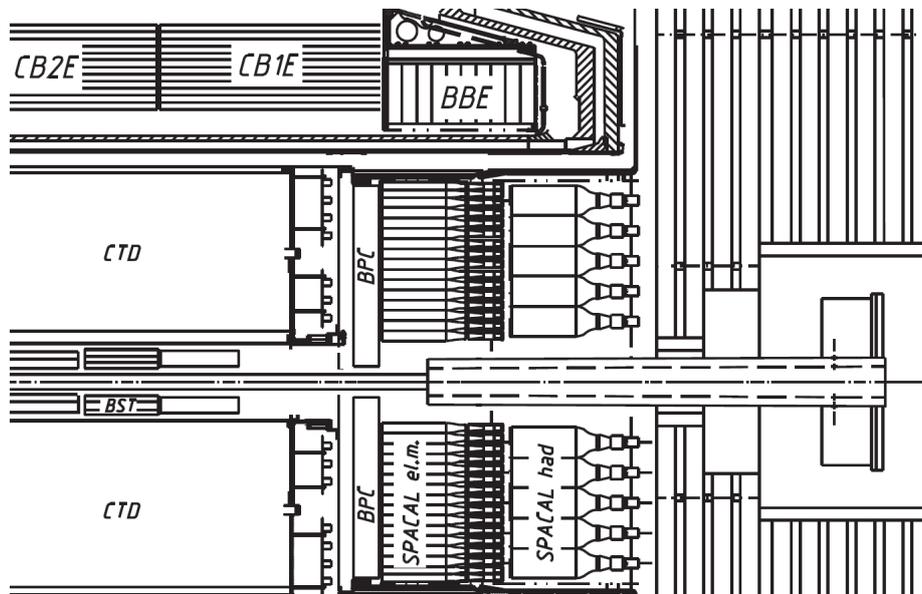
Figure 3.2.5: *Side view of the SpaCal calorimeter. Note its division in an electromagnetic and a hadronic part. In front of the calorimeter the track chamber BPC is mounted.*
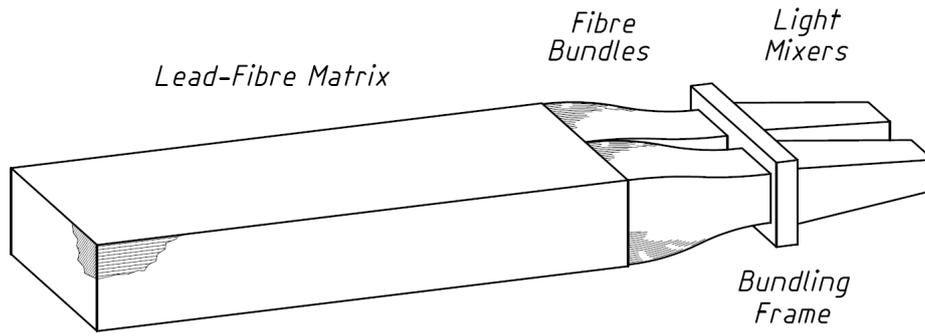
Figure 3.2.6: *A single cell of the electromagnetic SpaCal. It consists of 2340 longitudinally aligned scintillating fibers embedded in a lead matrix.*
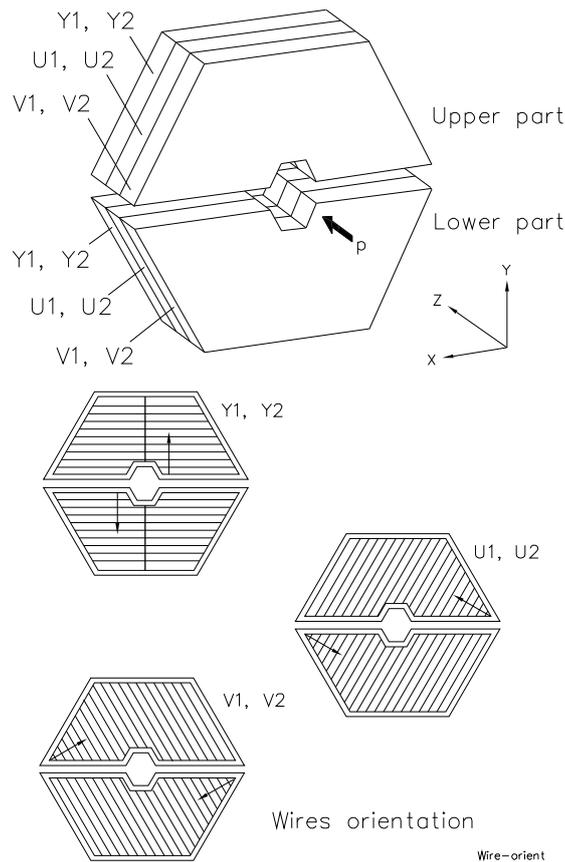
Figure 3.2.7: *Backward proportional chamber (BPC). It is mounted in front of the SpaCal. It features six layers with three azimuthal wire orientations. It supports the measurement of the spatial cluster position in the SpaCal.*

is achieved, which was determined with test beams at CERN and DESY. The spatial resolution of

$$\sigma_{xy} \simeq \frac{4 \text{ mm}}{\sqrt{E_{EL}/\text{GeV}}} \oplus 1 \text{ mm} \qquad (3.2.4)$$

is obtained for the reconstruction of the impact point [A$^+$96, N$^+$96].

The hadronic part of the SpaCal is made from 11.9 cm × 11.9 cm × 25 cm cells, comprising 1 mm fibers with a material ratio of 3.4 : 1. An energy resolution of

$$\sigma_{E_{HAD}}/E_{HAD} \simeq 29\% \qquad (3.2.5)$$

has been found with pion test beams [A$^+$96].

3.2.4   Luminosity System

The luminosity is measured with Bethe-Heitler events ($ep \to ep\gamma$), the cross section of which is known to high precision. The emitted photon is measured in the *Photon Detector (PD)*, a Cerenkov sampling calorimeter with tungsten absorbers [A$^+$97a]. Background mainly emerges from beam gas interactions, which are accounted for by the analysis of pilot bunches. Moreover, pile-up events from multiple interactions per bunch crossing occur. While the luminosity is monitored online in 10 second intervals, its precise value is determined offline, accounting for beam gas interactions, pile-up events and the efficiency and acceptance of the photon detector.

For the HERA II run phase, an independent luminosity determination is done with QED Compton events, searching for SpaCal clusters which are back to back in azimuthal angle ϕ. With this procedure, the systematic uncertainty of the luminosity measurement is reduced to ∼ 2%. At the same time, a small *QEDC Luminosity Correction* has to be applied to each run [SP].

3.3   The Trigger

The bunch crossing rate of 10.4 MHz largely exceeds the output rate of ∼ 50 Hz that can be achieved at H1. Therefore, a four-level trigger is used to suppress background and to downscale the frequency of abundant event types (*prescaling*). The trigger levels (L1,L2,L3 and L4/5) consecutively ascend in both precision and dead-time.

3.3.1   H1 Trigger Architecture

The L1 trigger level is pipelined and effectively dead-time free. It is composed of 128 *raw subtriggers*. If at least one of them is activated ("L1 keep"), the pipelines are stopped and dead time accumulates. The raw subtriggers are then logically combined of 256 *trigger elements (TE)*. The output rate of level L1 is ∼ 1 kHz.

The L2 trigger level is used for online verification on a time scale of 20 μs. A topological trigger (L2TT) [Biz97] providing 16 trigger elements is used, providing background reduction on the basis of topological event signatures. Moreover, a system with 13 neural networks (L2NN) [K$^+$97] is used.
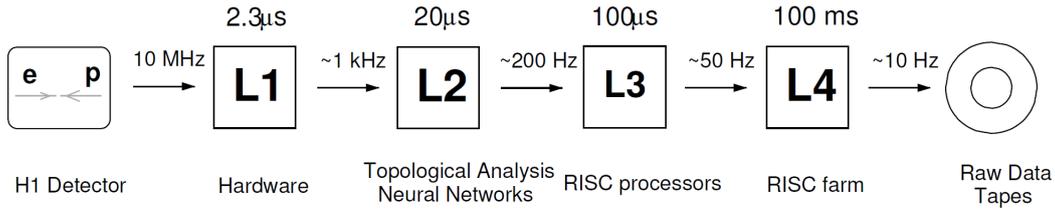
Figure 3.3.1: *H1 Trigger architecture featuring 4 trigger levels L1, L2, L3 and L4/L5. The L1 trigger is pipelined and dead time free.*
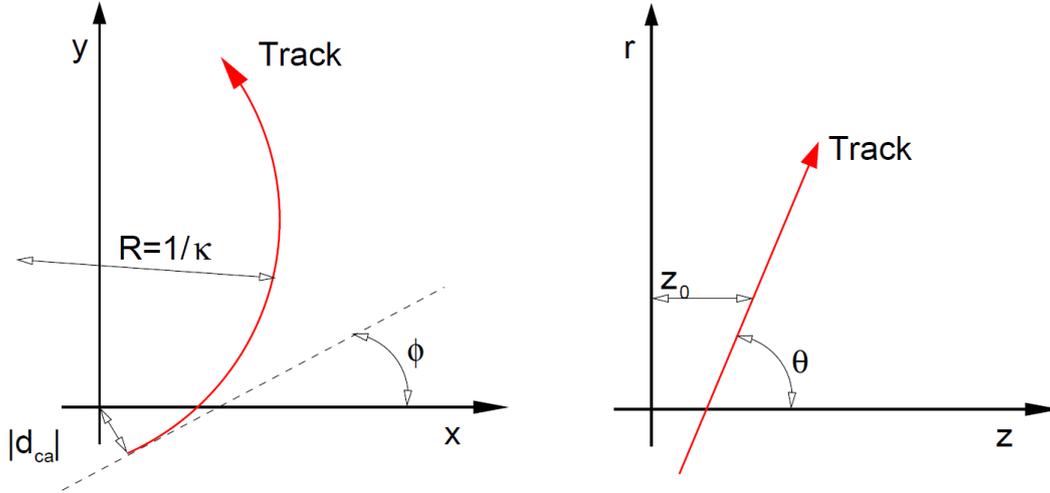
Figure 3.3.2: *The helical trajectory of a charged particle in the CTD with its 5 parameters κ, d_{CA}, ϕ_0, θ and z_0. Here, $\kappa = 1/R$, d_{CA} and ϕ_0 describe the track in the $r\phi$-plane, denoting respectively its inverse radius, its "distance of closest approach" to the origin and its azimuthal angle. In the rz-plane, the track is determined by the polar angle θ and the value z_0 of the point of closest approach to the z-axis.*

The L3 trigger is implemented since 2005. On a time scale of $\sim 100\mu s$ tracks from the FTT L3 were analyzed. It was mainly used to search for heavy quark decays [Nau, Jun07].

On the L4/L5 trigger level a complete event reconstruction is done using a dedicated PC farm. Further, the events are classified and downscaled according to their type. The output rate is ~ 10 Hz.

3.3.2   The S3 Trigger

In this analysis, the subtrigger S3 is used. It is sensitive to clusters in the electromagnetic part of the SpaCal. The cluster is required to have an energy $E_e^{uncalib}$ above $E_{tr} \equiv 10$ GeV and to be at least $R_{tr}^{uncalib} \equiv 30$ cm away from the beam axis:

$$E_e^{uncalib} > E_{tr} \equiv 10 \text{ GeV} \tag{3.3.1}$$

$$R_e^{uncalib} > R_{tr} \equiv 30 \text{ cm} \tag{3.3.2}$$

Since the trigger condition is applied at run time, no calibration is applied to the cluster energy and position, which is indicated by the superscript "*uncalib*". Note further, that
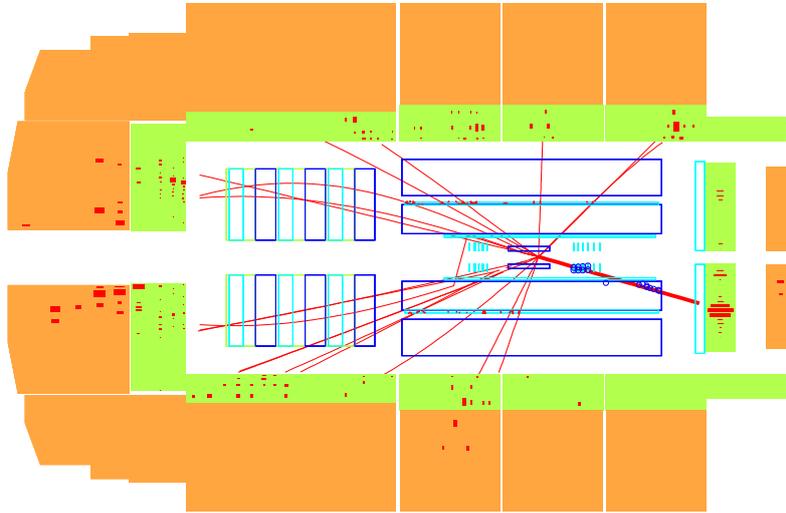
Figure 3.3.3: *A typical neutral current DIS event. Clearly visible are the scattered electron with its large SpaCal cluster and multiple reconstructed tracks from the hadronic final state.*

no prescaling is applied for S3. We will discuss the S3 efficiency in chapter 8.

3.4   Event Reconstruction

3.4.1   Track Reconstruction

Charged particles in a homogeneous magnetic field follow a helical trajectory. To account for multiple scattering, tracks are reconstructed using a *broken line fit* [Blo06], which allows the trajectory to be composed of two subtracks. This accounts for nuclear interactions with dead material, magnetic field inhomogeneities and energy loss. For the central region, the track finding algorithm starts from a few CJC hits (*track seed*) and successively adds hits from the CJC, COZ, CIP and silicon detectors. The resulting objects are called *Central Tracks.*

In figure 3.3.2 a helical trajectory is drawn with its 5 parameters κ, d_{CA}, ϕ_0, θ and z_0. Here, $\kappa = 1/R$, d_{CA} and ϕ_0 describe the track in the $r\phi$-plane, denoting respectively its inverse radius , its "distance of closest approach" to the origin and its azimuthal angle. In the rz-plane, the track is determined by the polar angle θ and the value z_0 of the point of closest approach to the z-axis.

A similar approach holds for tracks in the forward region of the detector (*Forward Tracks*). Moreover, *Combined Tracks* are defined via a matching algorithm using central and forward tracks.

3.4.2 z-Vertex Position

The longitudinal vertex position z_{Vtx} (*z-Vertex*) can be obtained using two different methods. Using the track detectors, it is obtained from essentially all tracks featuring a compatible value for d_{CA} in the $r\phi$-plane, see [Blo06]. Tracks from the Central Jet Chamber (CJC), the Forward Silicon Tracker (FST) and the Backward Silicon Tracker (BST) are used. A weighted average over the parameters z_0 is calculated. The other spatial coordinates of the vertex position are then derived from the beam geometry. Note, that from here on the event vertex obtained from this procedure is referred to as *CJC Vertex* z_{Vtx}^{Cjc}.

The next step is a refit of all tracks originating from the primary vertex using the CJC Vertex as a constraint. These objects are called *Vertex Fitted Tracks.*

The longitudinal vertex position can also be reconstructed using CIP hits from the scattered electron. This is done by a fit procedure developed by S. GLAZOV and S. PIEC [Gla06]. First, CIP hits along a line between the electron SpaCal position and the nominal vertex are searched for. Then, a linear fit in the rz-plane is done using the SpaCal cluster position as a constraint. The event vertex obtained from this procedure is referred to as *CIP Vertex* z_{Vtx}^{Cip}.

3.4.3 SpaCal Electron

To find electromagnetic particles in the SpaCal, a clustering algorithm is employed [Sch96]. The *cluster energy* is then simply the sum over the energy deposits in all contained cells i:

$$E_e := \sum_i E_i \tag{3.4.1}$$

Moreover, we define the *cluster position* as a weighted sum over the positions of all cells:

$$\vec{r} := \frac{\sum_i w_i \vec{r_i}}{\sum_i w_i} \tag{3.4.2}$$

The weight function is given by [A$^+$92]:

$$w_i = \max\left(0, w_0 + \ln \frac{E_i}{\sum_i E_i}\right) \tag{3.4.3}$$

By choosing $w_0 = 4.8$ an energy threshold is introduced.

The cluster size *ECRA* is defined in a similar manner [Gla], which is shown to be rather independent from the cluster position:

$$ECRA := \sqrt{\frac{\sum_i \tilde{w}_i \left(\vec{r_i} - \vec{r}\right)^2}{\sum_i \tilde{w}_i}} \tag{3.4.4}$$

Here, $\tilde{w}_i$ is defined analogously to (3.4.3). To account for imperfections in the shower simulation [GRP90] different thresholds are used for data ($\tilde{w}_0^{data} = 4.85$) and Monte Carlo ($\tilde{w}_0^{MC} = 5.05$).

This cluster size *ECRA* can be used as a discriminator to reduce hadronic background, which will be discussed in chapter 6.

3.5 Reconstruction of Kinematic Variables

At its core, this analysis is a counting experiment, in which all events are binned according to their inelasticity y and momentum exchange Q^2. The kinematic variables y and Q^2, however, cannot be measured directly, but have to be reconstructed from other variables, which are directly accessible trough the detector. There are different ways to achieve this, which are referred to as *reconstruction methods*.

In this analysis, the *electron method* is employed by default and the *double angle method* is used for calibration checks. Both techniques are described below. Other common methods at H1 are the *sigma method* and the *sigma electron method*. A review of the different reconstruction methods and a comparison of their strengths and weaknesses is given in [Spi92].

3.5.1 Electron Method

In this analysis, the electron method is used. The *electron method* obtains the kinematic invariants y_e and Q_e^2 from the *energy deposit E_e* and the *polar scattering angle θ_e* of the scattered electron. From section 2.1 we see, that for Born level events we have:

$$y \simeq y_e := 1 - \frac{E_e}{E_e^{beam}} \cdot \sin^2\left(\frac{\theta_e}{2}\right) \tag{3.5.1}$$

$$Q^2 \simeq Q_e^2 := 2E_e E_e^{beam} \cdot (1 - \cos\theta_e) \tag{3.5.2}$$

From analytical error propagation we obtain for the relative resolution:

$$\frac{\Delta y_e}{y_e} = \frac{1 - y_e}{y_e} \sqrt{\left(\frac{\Delta E_e}{E_e}\right)^2 + \left(\frac{\Delta\theta_e}{\tan\left(\frac{\theta_e}{2}\right)}\right)^2} \tag{3.5.3}$$

$$\frac{\Delta Q_e^2}{Q_e^2} = \sqrt{\left(\frac{\Delta E_e}{E_e}\right)^2 + \left(\tan\left(\frac{\theta_e}{2}\right) \cdot \Delta\theta_e\right)^2} \tag{3.5.4}$$

The advantage of this method is, that it makes use of the high SpaCal precision in E_e and θ_e, which is given by (3.2.3) and (3.2.4). All other reconstruction methods employ the hadronic final state (HFS), which is reconstructed with far less precision in both energy and direction. In [BB95] different reconstruction methods are compared for the H1 detector using the standard H1 detector simulation. They show, that for medium and high y (i.e. $y \gtrsim 0.1$) and low Q^2 (i.e. $7 \simeq Q^2$ GeV^2) the electron method features a better resolution in both Q^2 and x than the double angle and the sigma method. This makes the electron method especially suitable for an inclusive SpaCal analysis.

According to (3.5.3), the y_e-resolution depends strongly on y_e itself. Note, that the y_e-resolution is mainly limited by the energy resolution, since typically we have $\Delta\theta_e \sim 10^{-3}$. The Q_e^2-measurement degrades for $\theta_e \to \pi$ ("beampipe") due to the $\tan\left(\frac{\theta_e}{2}\right)$-factor in (3.5.4). Especially for high θ_e the resolution of θ_e becomes the limiting factor for the Q_e^2-resolution.

A draw-back of the electron method is its susceptibility to radiative effects. Consider an ISR event emitting a photon from the incoming electron line. Then, the

formulae (3.5.1) and (3.5.2) do not hold, and y and Q^2 will be wrongly reconstructed. A way to deal with this issue is described in subsection 5.1.4 and performed in chapter 11. It essentially introduces a bin wise correction to the cross section measurement, that is determined by a Monte Carlo simulation.

3.5.2   Double Angle Method

In contrast to the electron method, the *double angle method* utilizes the hadronic final state (HFS). Moreover, it is independent of the SpaCal energy measurement E_e and such can be used for calibration cross checks of the SpaCal calorimeter. Let the *hadronic angle* γ_h be given by:

$$\tan\left(\frac{1}{2}\gamma_h\right) := \frac{(E - p_z)_{HFS}}{(p_t)_{HFS}} \tag{3.5.5}$$

It can be viewed as the polar angle of the scattered parton. Then, we can define estimators for y, E and Q^2 via:

$$y \simeq y_{DA} := \frac{\tan\left(\frac{1}{2}\gamma_h\right)}{\tan\left(\frac{1}{2}\gamma_h\right) + \tan\left(\frac{1}{2}\theta_e\right)} \tag{3.5.6}$$

$$E \simeq E_{DA} := E_e^{beam} \cdot \frac{1 - y_{DA}}{\sin^2\left(\frac{\theta_e}{2}\right)} \tag{3.5.7}$$

$$Q^2 \simeq Q_{DA}^2 := 2E_{DA}E_e^{beam} \cdot (1 - \cos\theta_e) \tag{3.5.8}$$

We will make use of the double angle method in chapter 9.

3.6   SpaCal Calibration and Alignment

This analysis is essentially based on measurements of the SpaCal cluster energy E_e and position R_e. Therefore, we briefly review the calibration and alignment of the electromagnetic part of the SpaCal calorimeter.

3.6.1   SpaCal Energy Calibration

The energy calibration is done in multiple steps. During data taking, the performance of the photo multiplier tubes is monitored using pulsed LED signals and cross checked using cosmic muons, establishing a raw cluster energy measurement E_e^{raw} for each event. On analysis level, further calibration is done using the *double angle method*, which compares the reconstructed energy E_{DA} according to (3.5.7) with the raw cluster energy E_e^{raw}. From this comparison, a correction prescription is derived, which in turn is used to correct the cluster energy, $E_e^{raw} \rightarrow E_e$. The correction prescription uses both cell-specific calibration constants as well as functions of the cluster position [Col09].

The double angle method makes no reference to other calorimetric measurements. However, it evidently relies on the quality of the measurement of the angles θ_e and γ_h. Therefore, a special data subset around the kinematic peak ($E_e \sim E_{beam} \equiv 27.6$ GeV) with well-reconstructed electron and hadronic final state is choosen as calibration

sample, see [Col09]. The linearity of the energy measurement down to smaller electron energies is tested with $\pi^0 \to \gamma\gamma$ decays.

Since cross section measurements are performed by comparison with a Monte Carlo simulation, the focus of calibration tests is not on the absolute calibration scale, but rather on the level of agreement of between data and simulation. We will perform calibration tests in chapter 9.

3.6.2 SpaCal Alignment

The H1 coordinate system is defined by the positions of the CJC wires (*x*- and *y*-coordinates) and the COZ (*z*-coordinates). Therefore, all other detector components are aligned in respect to these two subdetectors.

The alignment of the SpaCal and the BPC is done using high energy electrons. The polar and azimuthal angle of both, SpaCal clusters and CTD tracks are compared, i.e. $\Delta\theta = \theta_e - \theta_{CTD}$ and $\Delta\phi = \phi_e - \phi_{CTD}$, see [Col09]. We will perform alignment tests in chapter 9.

Chapter 4

Unfolding

Every measurement in experimental physics is influenced by the limited resolution and the limited efficiency of the measuring device. Therefore, the measured distribution will be distorted in many ways; moreover, statistical correlations between the data points arise. However, our final goal is a cross section measurement, that is free from all such detector effects.

Different ways exists to treat these issues. The traditional way is a bin wise efficiency correction, which will be discussed in section 4.7. If the resolution of the detector is a limiting factor for the measurement precision, this method is not sufficient, since migration effects are not fully accounted for by this method.

The way unfolding addresses these issues is modelling the detector with all its resolution effects by a linear transformation $\boldsymbol{\mathcal{A}}$. This transformation maps a "true" distribution $\xi : \mathbb{R}^n \to \mathbb{R}, (\alpha_1, \ldots, \alpha_n) \to \xi(\alpha_1, \ldots, \alpha_n)$ of a set of n variables $\alpha_1, \ldots, \alpha_n$ onto a potentially distorted "measured" distribution $\boldsymbol{\mathcal{A}}[\xi] : \mathbb{R}^n \to \mathbb{R}, (\alpha_1, \ldots, \alpha_n) \to \boldsymbol{\mathcal{A}}[\xi](\alpha_1, \ldots, \alpha_n)$ of those n quantities:

$$\boldsymbol{\mathcal{A}} : \xi \to \boldsymbol{\mathcal{A}}[\xi]$$

With this transformation at hand, distortions in a measured distribution $\mathcal{N}(\alpha_1, \ldots, \alpha_n)$ can now be remedied by finding a distribution $\xi(\alpha_1, \ldots, \alpha_n)$ which satisfies:

$$\boldsymbol{\mathcal{A}}[\xi] = \mathcal{N} \tag{4.0.1}$$

If both the true and the measured distributions are given as discrete histograms $\vec{x}$ and $\boldsymbol{\mathcal{A}}[\vec{x}]$, then $\boldsymbol{\mathcal{A}}$ can be described by a matrix $\boldsymbol{A}$ (*response matrix*), such that we have $\boldsymbol{\mathcal{A}}[\vec{x}] = \boldsymbol{A}\vec{x}$. A measured distribution $\vec{N}$ can then be corrected by solving the following simple equation with respect to $\vec{x}$:

$$\vec{N} = \boldsymbol{A} \cdot \vec{x} \tag{4.0.2}$$

A typical difficulty of the unfolding technique is the rank deficiency of $\boldsymbol{A}$. It emerges due to the limited accuracy of the detector, which cannot discrimiate structures in the data below its resolution scale. Therefore, the solution $\vec{x}$ becomes ambiguous, hence (4.0.2) is under-constrained. This may result in large fluctuations in $\vec{x}$, large error bars and large bin to bin correlations.

To remedy this, a smooting prescription is imposed on the solution $\vec{x}$, which is denoted as *regularization*. At the price of potentially introducing a bias, the measurement is thus freed from those unphysical fluctuations. Of course, a reasonable

trade-off needs to be found. The determination of the right level of regularization is a key challenge of each unfolding procedure.

In this chapter we discuss some of the general aspects of unfolding, as far as they are needed for this analysis. We will introduce the basic definitions and the conceptual framework, while its application to the data will be forwarded to chapter 10. The reader who is familiar with unfolding might skip this chapter and come back to it when needed. For further reading on unfolding, see [Cow, Blo02, D'A95, VB98, Zec, Han00, HK96].

Note, that unless specified otherwise, we use the symbol Δ for the absolute error and the symbol δ for the relative error.

4.1 Detector effects

The way a detector affects the measurement is referenced to as *detector effects*.

Detector effects are studied using Monte Carlo simulations. In general, Monte Carlo events are produced in three steps, which are denoted as *generation, simulation* and *reconstruction*. During the generation phase, the physics properties of events are simulated including event type and kinematic of all outgoing jets and particles. This leads to *generator level distributions* of event specific variables. During the simulation phase, the interaction of the outcoming particles with the detector and the signals in the different detector components are simulated. From these simulated signals, tracks and clusters are reconstructed just as it is done for the data events. Distributions of reconstructed quantities are in this work referred to as *detector level distributions*. More information on the simulation process is given in chapter 7.

To study detector effects, one can compare distributions on generator level and detector level. Evidently, any quantitative determination depends strongly on the quality of the Monte Carlo simulation. We will come back to this issue in more detail in chapter 7 and 8.

There are two kinds of detector effects:

- *Efficiency effects*

- *Migration effects*

We will address both issues seperately.

4.1.1 Efficiency

In the context of this analysis multiple notions of efficiency have to be distinguished carefully.

We refer to the *reconstruction efficiency* ϵ_i^{rec} in bin i as the probability for an event originating from bin i to be measured *somewhere* in the detector (potentially even in an underflow or overflow bin). This probability can be estimated using a Monte Carlo simulation. Let N_{ji}^{MC} be the number of simulated events that have been generated in bin i and reconstructed in bin j. And let N_i^{gen} be the number of simulated events that have been generated in bin i. Then we have:

$$N_i^{gen} = \sum_j^{incl} N_{ji}^{MC} + N_i^{Lost} \tag{4.1.1}$$

Here, the sum runs over all detector level bins *including* overflow and underflow bins and N_i^{Lost} quantifies the events that have been generated in bin i and have *not* been reconstructed.

Then, the reconstruction efficiency ϵ_i^{rec} can be estimated by the following ratio $\hat{\epsilon}_i^{rec}$:

$$\epsilon_i^{rec} \simeq \hat{\epsilon}_i^{rec} := \frac{\sum_j^{incl} N_{ji}^{MC}}{N_i^{gen}} \tag{4.1.2}$$

Note, that the notion of the reconstruction efficiency gathers all efficiency effects related to the measurement. These are the sensitivity and acceptance of the detector, the efficiency of the reconstruction algorithm, the trigger efficiency and the efficiency of applied analysis level cuts.

Another reason for efficiency losses lies in the limited phase space region, which is under investigation (the *bin grid area*). An event that originates within the bin grid can be reconstructed outside the bin grid due to migration effects. We refer to the *bin grid efficiency* ϵ_i^{grid} in bin i as the probability for an event that originates from bin i *and* is measured in the detector, to be measured within the bin grid area. Again, this probability can be estimated using a Monte Carlo simulation.

Using the above nomenclature, the bin grid efficiency ϵ_i^{grid} can be estimated by the following ratio $\hat{\epsilon}_i^{grid}$:

$$\epsilon_i^{grid} \simeq \hat{\epsilon}_i^{grid} := \frac{\sum_j^{excl} N_{ji}^{MC}}{\sum_j^{incl} N_{ji}^{MC}} \tag{4.1.3}$$

Here, the sum in the numerator runs over all detector level bins *excluding* the overflow and underflow bins, while for the sum in the denominator these bins are *included*.

The product of reconstruction efficiency and bin grid efficiency will be called the *overall efficiency* ϵ_i and is estimated by the following expression:

$$\epsilon_i := \epsilon_i^{reco} \cdot \epsilon_i^{grid} \simeq \frac{\sum_j^{excl} N_{ji}^{MC}}{N_i^{gen}} \tag{4.1.4}$$

The absolute statistical uncertainty $\Delta\epsilon_i$ of this quantity can be calculated in a binomial manner:

$$\Delta\epsilon_i = \sqrt{\frac{\epsilon_i \cdot (1 - \epsilon_i)}{N_i^{gen}}} \tag{4.1.5}$$

Similar expressions hold for the reconstruction efficiency and the bin grid efficiency.

A widely used approximation for the overall efficiency is the fraction between the number of reconstructed events N_i^{rec} and the number of generated events N_i^{gen} in bin i. Since this quantity plays an important role in the context of bin-by-bin unfolding, we will refer to it as *bin-by-bin efficiency*:

$$\epsilon_i^{bbb} := \frac{N_i^{rec}}{N_i^{gen}} \tag{4.1.6}$$

Here, we use:

$$N_i^{rec} := \sum_k^{incl} N_{ik}^{MC} \tag{4.1.7}$$

Again, the sum runs over all generator level bins including overflow and underflow bins. It must be emphasized that - in spite of its name - this quantity is not an efficiency in the commonly used sense. In particular, it can be greater than 1, because the number of reconstructed events N_i^{rec} can in some cases exceed the number of generated events N_i^{gen} due to migration effects. Moreover, the formula (4.1.5) does not hold for the uncertainty of the bin-by-bin efficiency, as can easily be seen by inserting $\epsilon_i^{bbb} > 1$ into (4.1.5). Instead, by treating the simulated events N_{ij}^{MC} as POISSON-distributed, $\Delta N_{ij}^{MC} \simeq \sqrt{N_{ij}^{MC}}$, we obtain:

$$\Delta \epsilon_i^{bbb} = \frac{N_i^{rec}}{N_i^{gen}} \cdot \sqrt{\frac{1}{N_i^{gen}} - 2 \cdot \frac{N_{ii}^{MC}}{N_i^{gen} N_i^{rec}} + \frac{1}{N_i^{rec}}} \qquad (4.1.8)$$

However, one typically uses a slightly more conservative approximation, which effectively treats N_i^{gen} and N_i^{rec} as statistically independent:

$$\Delta \epsilon_i^{bbb} \simeq \frac{N_i^{rec}}{N_i^{gen}} \cdot \sqrt{\frac{1}{N_i^{gen}} + \frac{1}{N_i^{rec}}} \qquad (4.1.9)$$

The determination of all these efficiencies depends strongly on the quality of the Monte Carlo simulation. Therefore, the simulated efficiencies have to be individually cross checked with real and independent data samples. If discrepancies show up, these can be addressed by corrrecting the Monte Carlo on detector level with an appropriate weight function. We will come back to these issues in detail in chapter 8.

4.1.2   Migration

Events are not necessarily measured in the bins they originate from. Instead, they can be reconstructed in one of the neighbouring bins. This effect is called *migration*. There are two common reasons for migration, *detector resolution* and *non-linear detector effects*. Limitations in the detector resolution lead to a smearing of the measured distribution. A non-linear detector response will result in a systematic distortion of the measured distribution.

One quantifies migration effects in terms of *stability* and *purity*. The *stability* s_i in bin i refers to the probability for an event originating from bin i to be measured in bin i. Again, this quantity has to be determined with a Monte Carlo simulation. Using the nomenclature of subsection (4.1.1), this quantity can be estimated by the following expression:

$$s_i \simeq \hat{s}_i := \frac{N_{ii}^{MC}}{N_i^{gen}} = \frac{N_{ii}^{MC}}{\sum_k^{incl} N_{ki}^{MC} + N_i^{Lost}} \qquad (4.1.10)$$

The term *purity* p_i in bin i refers to the probability for an event being measured in bin i to originate in bin i. Using a Monte Carlo simulation this quantity can be estimated by the following expression:

$$p_i \simeq \hat{p}_i := \frac{N_{ii}^{MC}}{N_i^{rec}} = \frac{N_{ii}^{MC}}{\sum_k^{incl} N_{ik}^{MC}} \qquad (4.1.11)$$

Combination of (4.1.6), (4.1.10) and (4.1.11) shows a simple relation of stability, purity and bin-by-bin efficiency:

$$\epsilon_i^{bbb} = \frac{s_i}{p_i} \tag{4.1.12}$$

Both purity and stability depend strongly on the chosen bin-size. The choice of larger bins can help to reduce the impact of migration on the data. However, the chosen bin size puts a limit on the resolution of the measurement. Finding the right trade-off between those two counteracting goals can be complex and needs to be carefully adapted to the situation at hand. We will come back to this question in section 4.5.

Again, the determination of purity and stability depends strongly on the quality of the Monte Carlo simulation. Therefore, the calibration and the resolution of the detector has to be cross checked with independent data samples. This will be done in chapter 9.

4.2 Regularized Unfolding

Efficiency and resolution of the detector can be modelled by a *response matrix* $\boldsymbol{A}$. It relates a generated distribution with its simulated and reconstructed "measurement".

4.2.1 The Response Matrix

Let us choose a *generator level bin grid* $\mathcal{G}_{gen}$ with n bins $\mathcal{B}_i$ and a *detector level bin grid* $\mathcal{G}_{det}$ with m bins $\mathcal{B}_j^{det}$. Then any *generated distribution* can be depicted as a vector $\vec{N}^{gen}$ with n entries N_i^{gen} which hold the number of generated events in bin $\mathcal{B}_i$. Any *reconstructed Monte Carlo distribution* $\vec{N}^{rec}$ or any *measured data distribution* $\vec{N}^{data}$ is a vector with m entries, correspondingly.

Let A_{ji} be the probability of an event generated in bin $\mathcal{B}_i$ to be measured in Bin $\mathcal{B}_j^{det}$. Then, the *response matrix* $\boldsymbol{A}$ is defined as the $m \times n$-Matrix built from these probabilities:

$$\boldsymbol{A} := (A_{ji}) \tag{4.2.1}$$

The response matrix can be estimated by a Monte Carlo simulation. Using the nomenclature of section (4.1.1) we have the following expression:

$$A_{ji} \simeq \frac{N_{ji}^{MC}}{N_i^{gen}} \tag{4.2.2}$$

We refer to the matrix $\boldsymbol{N}^{MC}$ comprising the numbers N_{ji}^{MC} as *scaled response matrix*. Its entries N_{ji}^{MC} simply represent the number of events that have been generated in bin i and have been reconstructed in bin j:

$$\boldsymbol{N}^{MC} = \left(N_{ji}^{MC} \right) \tag{4.2.3}$$

Obviously, the generated distribution $\vec{N}^{gen}$ and its reconstructed counterpart $\vec{N}^{rec}$ are related by a simple linear equation:

$$\vec{N}^{rec} = \boldsymbol{A} \cdot \vec{N}^{gen} \tag{4.2.4}$$

The reponse matrix can be viewed as a mathematical model of the whole measurement process. All efficiency and migration effects are encoded in its elements. Note for example, that if the binning on generator and detector level is the same, a diagonal element A_{ii} of the response matrix corresponds to the stability of bin $\mathcal{B}_i$, see definition (4.1.10).

The basic idea of unfolding is now to utilize the response matrix to correct the impact of smearing on the measured data distribution, i. e. to solve the linear equation with respect to $\vec{x}$:

$$\vec{N}^{data} = \boldsymbol{A} \cdot \vec{x} \tag{4.2.5}$$

Here, the *unfolded distribution* $\vec{x}$ is a vector with n entries x_i which hold the corrected number of events in bin $\mathcal{B}_i$.

4.2.2   Unfolding as an inversion problem

In most cases equation (4.2.5) cannot be solved directly. The naive ansatz $\vec{x} = \boldsymbol{A}^{-1} \cdot \vec{N}^{data}$ is only defined, if $m = n$ and if $\det \boldsymbol{A} \neq 0$. Both conditions are usually not fulfilled. However, in most physics scenarios, one chooses a detector level bin grid $\mathcal{G}_{det}$ which is finer than the generator bin grid $\mathcal{G}_{gen}$, i. e. $m > n$. In this case, the equation system (4.2.5) is often over-determined and a "solution" can be found by minimizing the following χ^2-function with respect to $\vec{x}_0$:

$$\chi_A^2\left(\vec{x_0}\right) := \left(\vec{N}^{data} - \boldsymbol{A} \cdot \vec{x_0}\right)^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \left(\vec{N}^{data} - \boldsymbol{A} \cdot \vec{x_0}\right) \tag{4.2.6}$$

Here, the $\boldsymbol{COV}_{\vec{N}^{data}}^{-1}$ denotes the inverse covariance matrix of the measured data distribution $\vec{N}^{data}$ [Bloa].

It is a straightforward calculation to find the solution $\vec{x}_0$ to (4.2.6). First, we introduce the following abbreviation, in which we follow [Scha]:

$$\boldsymbol{E}_0^{-1} := \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \boldsymbol{A} \tag{4.2.7}$$

By setting $0 = \nabla_{\vec{\xi}} \chi_A^2\left(\vec{\xi}\right)\Big|_{\vec{\xi}=\vec{x}_0}$ one obtains:

$$\vec{x}_0 = \left[\boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \boldsymbol{A}\right]^{-1} \cdot \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \vec{N}^{data}$$

$$\vec{x}_0 = \boldsymbol{E}_0 \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \vec{N}^{data} \tag{4.2.8}$$

Using standard error propagation, we can directly compute the covariance matrix $\boldsymbol{COV}_{\vec{x}_0}$ for the unfolded distribution $\vec{x}_0$:

$$\boldsymbol{COV}_{\vec{x}_0} = \left[\boldsymbol{E}_0 \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1}\right] \cdot \boldsymbol{COV}_{\vec{N}^{data}} \cdot \left[\boldsymbol{E}_0 \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1}\right]^T$$

$$\boldsymbol{COV}_{\vec{x}_0} = \boldsymbol{E}_0 \tag{4.2.9}$$

Note, that if $m = n$ and $\det \boldsymbol{A} \neq 0$, the solution $\vec{x}_0$ to (4.2.6) is also the direct solution to (4.2.5). However, the solution $\vec{x}_0$ from (4.2.8) exists even under the much weaker condition, that $\boldsymbol{E}_0^{-1} \equiv \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \boldsymbol{A}$ is invertible, which is fulfilled in some physics cases.

More troublesome is the case, that $\boldsymbol{E}_0^{-1} \equiv \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \boldsymbol{A}$ is *not* invertible. In this much more frequent scenario, there is more than one solution to (4.2.6). From an experimental perspective, this can be easily understood with regard to the limited accuracy of the detector. Consider two distributions $\vec{x}^1$ and $\vec{x}^2$ with $\vec{x}^1 \neq \vec{x}^2$, which differ only on a scale much smaller than the dector resolution. This could for example be a narrow peak in one of the two distributions, which is two small to be resolved by the measuring device. Both distributions then yield the same detector response $\vec{N}^{data}$, i.e. we would have $\vec{N}^{data} = \boldsymbol{A}\vec{x}^1 = \boldsymbol{A}\vec{x}^2$ and both distributions minimize (4.2.6) equally well.

In a typical situation in experimental physics, the matrix $\boldsymbol{E}_0^{-1} \equiv \boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \boldsymbol{A}$ is invertible, but the determinant is close to zero, $\det\left(\boldsymbol{E}_0^{-1}\right) \simeq 0$. In this case, the covariance matrix $\boldsymbol{COV}_{\vec{x}_0} = \boldsymbol{E}_0$ will typically show large variances and correlations. This frequent case emerges, whenever we have enough measured data to *mathematically* determine the solution $\vec{x}_0$ to (4.2.6), but due to the limited detector resolution the data will not determine the solution *experimentally*, i.e. with reasonably small statistical uncertainties. Then, the solution $\vec{x}_0$ will typically show large statistical correlations between the entries. Moreover, the solution $\vec{x}_0$ will be determined by large fluctuations and large error bars, making it inconclusive.

4.2.3 Regularization

The core of the problem is the ambiquity in the solution of (4.2.6). This problem can only be overcome by making a choice! The method of *regularized unfolding* makes such a choice by imposing the condition of *smootheness* on the solution $\vec{x}$. This idea embraces the fact, that only sufficiently smooth solutions $\vec{x}$ to (4.2.6) are of interest, since any substructure below the resolution of the detector cannot be resolved.

We ensure the smoothness of the solution $\vec{x}$ by introducing a *penalty term* $\chi_L^2(\vec{x})$, which will be added to the function (4.2.6), see [Bloa]. Thus, solutions with weak fluctuations are favored in the minimization process:

$$\chi_L^2(\vec{x}) := \mathcal{C}(\vec{x} - \vec{x}^{Theory}) \tag{4.2.10}$$

$$\chi^2(\vec{x}) := \chi_A^2(\vec{x}) + \tau^2 \cdot \chi_L^2(\vec{x}) \tag{4.2.11}$$

Here, $\mathcal{C}(\vec{x})$ is a positive-valued function which measures the amount of fluctuations of the solution $\vec{x}$ (*curvature measure*). The parameter τ is a number greater or equal than zero (*regularization parameter*). Note, that in the definition of the penalty term, the curvature appears in reference to the difference $\vec{x} - \vec{x}^{Theory}$ to some theoretically expected distribution $\vec{x}^{Theory}$.

One of the most difficult and controversial issues regarding unfolding is the choice of the regularization parameter. An array of methods has been developed to find the right trade-off between smoothness on the one hand and safeguarding the measured data against bias on the other. Some of these methods have been tested in this analysis.

Moreover, there are multiple possibilities to define the curvature measure $\mathcal{C}(\vec{x})$, some of which will be discussed in section 4.3. They can be viewed as quadratic forms on the vector space of possible solutions $\vec{x}$:

$$\vec{x} \to \mathcal{C}(\vec{x}) = \vec{x}^T \boldsymbol{L}^2 \vec{x} \tag{4.2.12}$$

Here, $\boldsymbol{L}^2$ is a symmetric, positive semi-definite $n \times n$-Matrix, where n is the number of generator level bins. With this notation, (4.2.10) becomes:

$$\chi_L^2(\vec{x}) = \left(\vec{x} - \vec{x}^{Theory}\right)^T \boldsymbol{L}^2 \left(\vec{x} - \vec{x}^{Theory}\right) \tag{4.2.13}$$

To minimize (4.2.11), we define [Scha]:

$$\boldsymbol{E}^{-1} := \boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \boldsymbol{A} + \tau^2 \cdot \boldsymbol{L}^2 \tag{4.2.14}$$

By setting $0 = \nabla_{\vec{\xi}} \chi_A^2\left(\vec{\xi}\right)\Big|_{\vec{\xi}=\vec{x}}$, we obtain for the optimum $\vec{x}$:

$$\vec{x} = \left[\boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \boldsymbol{A} + \tau^2 \boldsymbol{L}^2\right]^{-1} \cdot \left[\boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \vec{N}^{data} + \tau^2 \boldsymbol{L}^2 \cdot x^{Theory}\right]$$

$$\vec{x} = \boldsymbol{E} \cdot \left[\boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \vec{N}^{data} + \tau^2 \boldsymbol{L}^2 \vec{x}^{Theory}\right] \tag{4.2.15}$$

Using standard error propagation, we can directly compute the covariance matrix $\boldsymbol{COV}_{\vec{x}}$ for the unfolded distribution $\vec{x}$:

$$\boldsymbol{COV}_{\vec{x}} = \left[\boldsymbol{E}\boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1}\right] \cdot \boldsymbol{COV}_{\vec{N}^{data}} \cdot \left[\boldsymbol{E}\boldsymbol{A}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1}\right]^T$$

$$\boldsymbol{COV}_{\vec{x}} = \boldsymbol{E} \tag{4.2.16}$$

4.3   The Curvature Measure

In the context of this section we discuss three different types of curvature measures, size regularization, derivational regularization and curvature regularization. These methods are implemented in the `Root`-package `TUnfold`, which has been written by S. SCHMITT [Schb].

4.3.1   Notions of Curvature

The most simple measure of curvature studied in this analysis is referred to as *size regularization $\mathcal{C}_{size}(\vec{x})$*. Here, the squared Euklidian norm of the vector $\vec{x}$ is studied:

$$\mathcal{C}_{size}(\vec{x}) := \sum_{i=1}^{n} x_i^2 = \vec{x}^T \boldsymbol{L}_{size}^2 \vec{x} \tag{4.3.1}$$

When applied to the minimization of (4.2.11), the solution $\vec{x}$ is forced to be close the theoretical expectation $\vec{x}^{Theory}$, thus statistical fluctuations are reduced.

The matrix $\boldsymbol{L}_{size}^2$ is diagonal and has unit trace elements. The key advantages of this ansatz are simplicity and the fact that it has been well studied in literature. Moreover, it can easily be applied to more dimensional unfolding problems. An obvious problem of the method is its tendency to bias the solution $\vec{x}$ towards the theoretical expectation $\vec{x}^{Theory}$. Note in particular, that this bias not only affects the shape, but also the normalisation of $\vec{x}$.

An alternative measure of curvature is called *derivational regularization $\mathcal{C}_{deriv}(\vec{x})$*. For a one dimensional problem, we define:

$$\mathcal{C}_{deriv,1D}(\vec{x}) := \sum_{i=1}^{n-1} \left(x_{i+1} - x_i\right)^2 = \vec{x}^T \boldsymbol{L}_{deriv,1D}^2 \vec{x} \tag{4.3.2}$$

Thus, for the matrix $\boldsymbol{L}_{deriv,1D}^2$ we obtain:

$$\boldsymbol{L}_{deriv,1D}^2 = \boldsymbol{L}_D \equiv \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \tag{4.3.3}$$

If the bin size is roughly constant throughout the bin grid, this measure can be viewed as a "discrete first derivative" of $\vec{x}$. Therefore, the minimization process favours solutions with small fluctuations while preserving the global normalisation at the same time.

The third measure of curvature is called *curvature regularization* $\mathcal{C}_{curve}(\vec{x})$. For a one dimensional problem, we define:

$$\mathcal{C}_{curve,1D}(\vec{x}) := \sum_{i=1}^{n-2} \left(x_{i+2} - 2 \cdot x_{i+1} + x_i \right)^2 = \vec{x}^T \boldsymbol{L}_{curve,1D}^2 \vec{x} \tag{4.3.4}$$

Thus, for the matrix $\boldsymbol{L}_{curve,1D}^2$ we obtain:

$$\boldsymbol{L}_{curve,1D}^2 = \boldsymbol{L}_C \equiv \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix} \tag{4.3.5}$$

If the bin size is roughly constant throughout the bin grid, this measure can be viewed as a "discrete second derivative" of $\vec{x}$. Therefore, the minimization process favours solutions with small fluctuations while the global normalization and the slope of the distribution are preserved.

Similar definitions can be found for the two dimensional case. Let the generator level bin grid $\mathcal{G}_{gen}$ have n_1 bins in the first and n_2 bins in the second direction. Then, $\vec{x}$ is a $n_1 \times n_2$-matrix with entries x_{i_1,i_2}, each representing the number of events in one of the bins $\mathcal{B}_{i_1,i_2}$. Moreover, the symbol $\boldsymbol{L}_{deriv,2D}^2$ denotes a four-dimensional tensor. However, we can still formulate the problem in the familiar framework of vectors and matrices by introducing an index mapping, which will be referred to as *bin numbering scheme*. This is done by allocating each bin $\mathcal{B}_{i_1,i_2}$ a unique bin number $i = \nu_{2D}(i_1, i_2)$ via the prescription:

$$\nu_{2D} : \{1, \ldots, n_1\} \times \{1, \ldots, n_2\} \to \{1, \ldots, n_1 \cdot n_2\}, (i_1, i_2) \to \nu_{2D}(i_1, i_2) \tag{4.3.6}$$

The choice of such a numering scheme is of course purely conventional, we choose a lexical mode:

$$\nu_{2D}(i_1, i_2) := i_1 \cdot (n_2 - 1) + i_2 \tag{4.3.7}$$

Unless specified otherwise, this bin numbering scheme will be used implicitly throughout this analysis.

We now write down a definition for the "discrete first derivative":

$$\mathcal{C}^{2D}_{deriv}(\vec{x}) := \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2-1} (x_{i_1,i_2+1} - x_{i_1,i_2})^2 + \sum_{i_1=1}^{n_1-1} \sum_{i_2=1}^{n_2} (x_{i_1+1,i_2} - x_{i_1,i_2})^2 = \vec{x}^T \boldsymbol{L}^2_{deriv,2D} \vec{x}$$
(4.3.8)

If the bin size is roughly constant throughout the bin grid, this measure can be thought of as the Euklidian norm of the "gradient" of the distribution $\vec{x}$. Using (4.3.7), we obtain for $\boldsymbol{L}^2_{deriv,2D}$ the following block matrix structure:

$$\boldsymbol{L}^2_{deriv,2D} = \begin{pmatrix} \boldsymbol{L}_D + \boldsymbol{I} & -\boldsymbol{I} & & & \\ -\boldsymbol{I} & \boldsymbol{L}_D + 2 \cdot \boldsymbol{I} & -\boldsymbol{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\boldsymbol{I} & \boldsymbol{L}_D + 2 \cdot \boldsymbol{I} & -\boldsymbol{I} \\ & & & -\boldsymbol{I} & \boldsymbol{L}_D + 2 \cdot \boldsymbol{I} \end{pmatrix}$$
(4.3.9)

Here, the symbol $\boldsymbol{L}_D$ is given by (4.3.3) and $\boldsymbol{I}$ denotes the unit matrix in n_1 dimensions.

Persuing the idea of a "discrete second derivative", we also define:

$$\mathcal{C}^{2D}_{curve}(\vec{x}) :=$$

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2-2} (x_{i_1,i_2+2} - 2 \cdot x_{i_1,i_2+1} + x_{i_1,i_2})^2 + \sum_{i_1=1}^{n_1-2} \sum_{i_2=1}^{n_2} (x_{i_1+2,i_2} - 2 \cdot x_{i_1+1,i_2} + x_{i_1,i_2})^2$$

$$\mathcal{C}^{2D}_{curve}(\vec{x}) = \vec{x}^T \boldsymbol{L}^2_{deriv,2D} \vec{x}$$
(4.3.10)

If the bin size is roughly constant throughout the bin grid, this measure can be thought of as the "Laplacian" of the distribution $\vec{x}$. Using (4.3.7), we obtain for $\boldsymbol{L}^2_{curve,2D}$ the following block matrix structure:

$$\boldsymbol{L}^2_{curve,2D} =$$
(4.3.11)

$$\begin{pmatrix} \boldsymbol{L}_C + \boldsymbol{I} & -2 \cdot \boldsymbol{I} & \boldsymbol{I} & & & \\ -2 \cdot \boldsymbol{I} & \boldsymbol{L}_C + 5 \cdot \boldsymbol{I} & -4 \cdot \boldsymbol{I} & \boldsymbol{I} & & \\ \boldsymbol{I} & -4 \cdot \boldsymbol{I} & \boldsymbol{L}_C + 6 \cdot \boldsymbol{I} & -4 \cdot \boldsymbol{I} & \boldsymbol{I} & \\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ & \boldsymbol{I} & -4 \cdot \boldsymbol{I} & \boldsymbol{L}_C + 6 \cdot \boldsymbol{I} & -4 \cdot \boldsymbol{I} & \boldsymbol{I} \\ & & \boldsymbol{I} & -4 & \boldsymbol{L}_C + 5 \cdot \boldsymbol{I} & -2 \cdot \boldsymbol{I} \\ & & & \boldsymbol{I} & -2 \cdot \boldsymbol{I} & \boldsymbol{L}_C + \boldsymbol{I} \end{pmatrix}$$

Here, the symbol $\boldsymbol{L}_C$ is given by (4.3.5) and $\boldsymbol{I}$ denotes the unit matrix in n_1 dimensions.

For the measurement of any differential distribution, the preservation of normalization and slope seems to be a critical requirement. This is particularly true for the measurement of a double differential ep cross section, which is dominated by large slopes along the kinematical variables Q^2 and y. On the other hand, statistical fluctuations are expected to appear mainly in the second derivative of the solution $\vec{x}$. This is due to the nature of migration, which leads to strong negative correlations between neighbouring data points x_i. In the light of all this, the curvature regularization seems to be the most adequate method and has indeed been chosen for this analysis.

4.3.2   Uniformity Conditions

While the usage of "discrete derivatives" appears to be far better motivated than simple size regularization, it can introduce various problems. The first one is a regional scaling problem. Consider a measurement suffering from migration effects which vary strongly between different regions of the phase space. The significance of a "peak" in the data will then depend on the phase space region; it can be of physical nature in areas of low migration and can be of purely statistical nature in highly smeared regions. However, the above measures do not account for that. The core of the problem is, that $\mathcal{C}_{deriv}(\vec{x})$ and $\mathcal{C}_{curve}(\vec{x})$ contain no information about the detector resolution $\Lambda(i)$ in bin $\mathcal{B}_i$, none about the chosen bin size $\mathrm{Vol}(i)$ and therefore none about the actual migration effects.

The second problem is a directional scaling problem; it is specific to more dimensional scenarios. Consider a two dimensional measurement suffering from migration effects which are very different between the two variables under investigation. Whether derivational changes in the solution $\vec{x}$ are of statistical or physical nature, depends on the direction. Again, the measures $\mathcal{C}_{deriv}(\vec{x})$ and $\mathcal{C}_{curve}(\vec{x})$ cannot account for that, since they lack information about the migration effects.

Both issues could in principle be addressed by introducing *scaling functions $Sc(i)$*, which encode the migration effects as a function of the bin number i, i.e. the phase space region. With these, the definitions (4.3.2) and (4.3.4) can be extended:

$$\mathcal{C}_{deriv,1D}^{scaled}(\vec{x}) := \sum_{i=1}^{n-1} Sc^{deriv}(i) \cdot (x_{i+1} - x_i)^2$$

$$\mathcal{C}_{curve,1D}^{scaled}(\vec{x}) := \sum_{i=1}^{n-2} Sc^{curve}(i) \cdot (x_{i+2} - 2 \cdot x_i + x_i)^2$$

For the two dimensional case we need two scaling functions $Sc_1(i_1, i_2)$ and $Sc_2(i_1, i_2)$ which are both funtions of both variables. The extended definitions of (4.3.8) and (4.3.10) are:

$$\mathcal{C}_{deriv,2D}^{scaled}(\vec{x}) :=$$

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2-1} Sc_1^{deriv}(i_1, i_2) \cdot (x_{i_1,i_2+1} - x_{i_1,i_2})^2 +$$

$$\sum_{i_1=1}^{n_1-1} \sum_{i_2=1}^{n_2} Sc_2^{deriv}(i_1, i_2) \cdot (x_{i_1+1,i_2} - x_{i_1,i_2})^2$$

$$\mathcal{C}_{curve,2D}^{scaled}(\vec{x}) :=$$

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2-2} Sc_1^{curve}(i_1, i_2) \cdot (x_{i_1,i_2+2} - 2 \cdot x_{i_1,i_2+1} + x_{i_1,i_2})^2 +$$

$$\sum_{i_1=1}^{n_1-2} \sum_{i_2=1}^{n_2} Sc_2^{curve}(i_1, i_2) \cdot (x_{i_1+2,i_2} - 2 \cdot x_{i_1+1,i_2} + x_{i_1,i_2})^2$$

While the definition of such scaling functions seems to be a highly non-trivial task, in many physics scenarios they turn out to be surprisingly simple. Consider, that the choice of the generator level bin grid $\mathcal{G}_{gen}$ is a fundamental decision of every analysis, which is driven by the two counteracting goals of a high number of data points on

the one hand and sufficient decorrelation of the data points on the other. Therefore, the bin size will often be chosen proportional to the detector resolution. For a two dimensional scenario, we would have:

$$\text{Length}^1 \left(\mathcal{B}_{i_1, i_2} \right) \simeq const \cdot \Lambda_1 \left(i_1, i_2 \right) \tag{4.3.12}$$

$$\text{Length}^2 \left(\mathcal{B}_{i_1, i_2} \right) \simeq const \cdot \Lambda_2 \left(i_1, i_2 \right) \tag{4.3.13}$$

Here, $\text{Length}^\mu \left(\mathcal{B}_{i_1, i_2} \right)$ denotes the length of the rectangular bin $\mathcal{B}_{i_1, i_2}$ in the μ-th direction and $\Lambda_\mu \left(i, j \right)$ measures the detector resolution in the μ-th direction. In this case, the level of migration will be roughly constant throughout the bin grid. In the context of this analysis, we speak of *regional conformity*:

$$s_{i_1, i_2} \simeq s_{l_1, l_2} \forall i_1, i_2, l_1, l_2 \tag{4.3.14}$$

Here, $s_{i_1 i_2}$ denotes the stability in bin $\mathcal{B}_{i_1, i_2}$. Moreover, the stability will not differ between the two directions of the bin grid, which we refer to as *directional conformity*:

$$s_{i_1}^1 \simeq s_{i_1}^2 \simeq s_{i_2}^1 \simeq s_{i_2}^2 \forall i_1, i_2 \tag{4.3.15}$$

Here, $s_{i_\lambda}^\mu$ denotes the stability in the μ-th direction at the bin i_λ in the direction λ.

If regional and directional conformity is given, we can set:

$$Sc_1(i_1, i_2) = Sc_2(i_1, i_2) = 1 \forall i_1, i_2 \tag{4.3.16}$$

The bin grid $\mathcal{G}_{gen}$ should be chosen such, that (4.3.14) and (4.3.14) are approximately fulfilled. This means, that the bin sizes have been choosen according to the resolution in both directions.

4.3.3   Bin Wise Measures of Curvature

The curvature measures discussed above can be defined in a bin wise manner [Schb]. This will be helpful especially in two dimensions, when we need to extend the curvature measures of subsection 4.3.1 to non-rectangular bin grids. Note, that in this subsection, we implicitly use the bin numbering scheme (4.3.7) for the two dimensional case, even if not explicitly mentioned.

We start with the size regularization for the one dimensional case. We introduce a positive semi-definite quadratic form $\mathcal{K}_{size}^i(\vec{x})$ for every bin $\mathcal{B}_i$ with the following simple structure:

$$\mathcal{K}_{size}^i(\vec{x}) := x_i^2 = \vec{x}^T \boldsymbol{K}_{size}^i \vec{x} \tag{4.3.17}$$

$$\left(\boldsymbol{K}_{size}^i \right)_{lk} = \delta_{il} \cdot \delta_{ik} \tag{4.3.18}$$

Now we define the curvature measure of size regularization $\mathcal{C}_{size}$ as the sum of the bin wise measures $\mathcal{K}_{size,1D}^i$:

$$\mathcal{C}_{size} := \sum_i \mathcal{K}_{size}^i \tag{4.3.19}$$

Here, the sum runs over all bins $\mathcal{B}_i$ that enter the unfolding process. By comparison with (4.3.1), we have redefined the curvature measure of size regularization from subsection 4.3.1.

For derivational regularization, we introduce a positive semi-definite quadratic form $\mathcal{K}_{deriv}^{(i,j)}(\vec{x})$ for every *neighbouring pair*[1] $(\mathcal{B}_i, \mathcal{B}_j)$ of bins with the following structure:

$$\mathcal{K}_{deriv}^{(i,j)}(\vec{x}) := (x_i - x_j)^2 = \vec{x}^T \boldsymbol{K}_{deriv}^{(i,j)} \vec{x} \tag{4.3.20}$$

$$\left(\boldsymbol{K}_{size}^{(i,j)} \right)_{lk} = \delta_{li} \cdot \delta_{ki} - \delta_{li} \cdot \delta_{kj} - \delta_{lj} \cdot \delta_{ki} + \delta_{lj} \cdot \delta_{kj} \tag{4.3.21}$$

Now we define the measure of derivational regularization $\mathcal{C}_{deriv}$ as the sum of these bin wise measures $\mathcal{K}_{deriv}^{(i,j)}$:

$$\mathcal{C}_{deriv} := \sum_{(i,j)} \mathcal{K}_{deriv}^{(i,j)} \tag{4.3.22}$$

Here, the sum runs over all neighbouring pairs $(\mathcal{B}_i, \mathcal{B}_j)$ of bins that enter the unfolding process. By comparison with (4.3.2) and (4.3.8), we have redefined the curvature measure of derivational regularization from subsection 4.3.1.

For curvature regularization, we introduce a positive semi-definite quadratic form $\mathcal{K}_{curve}^{(h,i,j)}(\vec{x})$ for every *neighbouring triple*[1] $(\mathcal{B}_h, \mathcal{B}_i, \mathcal{B}_j)$ of bins with the following structure:

$$\mathcal{K}_{curve}^{(h,i,j)}(\vec{x}) := (x_h - 2 \cdot x_i - x_j)^2 = \vec{x}^T \boldsymbol{K}_{curve}^{(h,i,j)} \vec{x} \tag{4.3.23}$$

$$\left(\boldsymbol{K}_{curve}^{(h,i,j)} \right)_{lk} = \delta_{lh} \cdot \delta_{kh} - 2 \cdot \delta_{lh} \cdot \delta_{ki} + \delta_{lh} \cdot \delta_{kj}$$
$$- 2 \cdot \delta_{li} \cdot \delta_{kh} + 4 \cdot \delta_{li} \cdot \delta_{ki} - 2 \cdot \delta_{li} \cdot \delta_{kj} \tag{4.3.24}$$
$$+ \delta_{lj} \cdot \delta_{kh} - 2 \cdot \delta_{lj} \cdot \delta_{ki} + \delta_{lj} \cdot \delta_{kj}$$

Now we define the curvature measure of curvature regularization $\mathcal{C}_{curve}$ as the sum of the bin wise measures $\mathcal{K}_{curve}^{(h,i,j)}$:

$$\mathcal{C}_{curve} := \sum_{(h,i,j)} \mathcal{K}_{curve}^{(h,i,j)} \tag{4.3.25}$$

Here, the sum runs over all neighbouring triples $(\mathcal{B}_h, \mathcal{B}_i, \mathcal{B}_j)$ of bins that enter the unfolding process. By comparison with (4.3.4) and (4.3.10), we have redefined the curvature measure of derivational regularization from subsection 4.3.1.

The advantage of this nomenclature becomes instantly apparent, when a curvature measure needs to be defined for a non-rectangular two dimensional bin grid. In these cases, the "property of neighbourhood" of a pair $(\mathcal{B}_i, \mathcal{B}_j)$ or triple $(\mathcal{B}_h, \mathcal{B}_i, \mathcal{B}_j)$ of bins needs to be checked individually and the definitions (4.3.22) and (4.3.25) can be adapted, accordingly. Definitions (4.3.8) and (4.3.10) fail to account for such cases.

Like the definitions from subsection 4.3.1, the bin wise curvature measures can be viewed as the discretization of continous curvature measures. For the one dimensional case, consider the distributions

$$K_\xi : \mathcal{C}_2(\mathbb{R}) \to \mathbb{R}, f \to \mathcal{K}_\xi(f) := \left. [f(x)]^2 \right|_{x=\xi}$$

$$\mathcal{K}_\xi' : \mathcal{C}_2(\mathbb{R}) \to \mathbb{R}, f \to \mathcal{K}_\xi'(f) := \left. \left[\frac{\partial}{\partial x} f(x) \right]^2 \right|_{x=\xi}$$

---

[1]Whether two or more bins fulfill the property of neighbourhood needs to be determined according to the geometry of the problem.

$$\mathcal{K}_\xi'' : \mathcal{C}_2(\mathbb{R}) \to \mathbb{R}, f \to \mathcal{K}_\xi''(f) := \left[\frac{\partial^2}{\partial x^2} f(x) \right]^2 \Bigg|_{x=\xi}$$

By discretization, we directly recover the measures $\mathcal{K}_{size}^i$, $\mathcal{K}_{deriv}^{(i,i+1)}$ and $\mathcal{K}_{curve}^{(i,i+1,i+2)}$ according to (4.3.17), (4.3.23) and (4.3.23).

4.4   Adjustment of Regularization

One of the most difficult issues of the unfolding process is the determination of the regularization parameter τ, which determines the level of smoothing in the minimization process (4.2.11). In this context, three different methods will be discussed which each motivate their own choice of τ.

4.4.1   The L-curve Method

The *L-curve method* is a graphical approach to adjust the level of regularization [Han00]. The `Root`-package `TUnfold` provides an implementation, see also the documentation there [Schb].

We start from a parametric plot of the two terms $\chi_A^2(\vec{x}(\tau))$ and $\chi_L^2(\vec{x}(\tau))$ in (4.2.11), which are drawn as a logarithmic function of τ:

$$\mathcal{L} : \tau \to \begin{cases} \mathcal{L}_x(\tau) := \log \chi_A^2(\vec{x}(\tau)) \\ \mathcal{L}_y(\tau) := \log \chi_L^2(\vec{x}(\tau)) \end{cases} \tag{4.4.1}$$

Here, $\vec{x}(\tau)$ is given by (4.2.15). Note, that $\mathcal{L}_x(\tau)$ measures the *data compatibility* of the solution $\vec{x}(\tau)$, while $\mathcal{L}_y(\tau)$ measures its *smoothness*. Therefore, $\mathcal{L}_x(\tau)$ rises monotonously with τ, while $\mathcal{L}_y(\tau)$ monotonously decreases. For small τ, $\mathcal{L}_y(\tau)$ typically decends fast, while $\mathcal{L}_x(\tau)$ stays constant. For large τ, however, a decrease of $\mathcal{L}_y(\tau)$ always comes at the price of an increase in $\mathcal{L}_x(\tau)$. As a result of all this, the plot of the function $\mathcal{L}$ will typically have the shape of an "L", showing a more or less distinct *kink* on its course. For an example, see figure 10.2.1 in chapter 10. The position of the kink is regarded the best compromise between data compatability and smoothness.

To pin down the L curve kink mathematically, we search for a value τ_L such that the function $\mathcal{L}$ has maximum curvature at that point. The program `TUnfold` achieves this by a scanning procedure over m scan points $\tau_1, \ldots, \tau_i, \ldots, \tau_m$. First, the values of $\mathcal{L}_x(\tau_i)$ and $\mathcal{L}_y(\tau_i)$ are determined for all i; then, both lists of pairs $(\tau_i, \mathcal{L}_x(\tau_i))$ and $(\tau_i, \mathcal{L}_y(\tau_i))$ are fitted by spline functions of third order $\mathcal{L}_x^{spline}(\tau)$ and $\mathcal{L}_y^{spline}(\tau)$. By construction we then have $\mathcal{L}_x^{spline}(\tau) \simeq \mathcal{L}_x(\tau)$ and $\mathcal{L}_y^{spline}(\tau) \simeq \mathcal{L}_y(\tau)$. Now, for each point τ_i the following curvature measure $\mathcal{C}_\mathcal{L}(\tau_i)$ is compared:

$$\mathcal{C}_\mathcal{L}(\tau_i) := \frac{\dot{\mathcal{L}}_x^{spline}(\tau) \cdot \ddot{\mathcal{L}}_y^{spline}(\tau) - \dot{\mathcal{L}}_y^{spline}(\tau) \cdot \ddot{\mathcal{L}}_x^{spline}(\tau)}{\sqrt{\left(\dot{\mathcal{L}}_x^{spline}(\tau)\right)^2 + \left(\dot{\mathcal{L}}_y^{spline}(\tau)\right)^2}^3} \Bigg|_{\tau=\tau_i} \tag{4.4.2}$$

Here, the symbols $\dot{\mathcal{L}}_z^{spline}(\tau)$ and $\ddot{\mathcal{L}}_z^{spline}(\tau)$ denote the first and second derivatives of the spline functions with respect to τ with $z = x, y$. For τ_L we choose the value that maximizes $\mathcal{C}_\mathcal{L}(\tau_i)$.

One of the advantages of the L curve method is, that it is intuitive, well-motivated and simple to use. It works equally well with either size regularization, derivational regularization or curvature regularization.

It must be emphasized tough, that the L curve method intends to simply find a compromise between data compatability and smoothness. It implicitly separates small-scale structures (high χ_L^2) in the solution $\vec{x}$ from large-scale structures (small χ_L^2) and damps the small-scale structures away. However, even small-scale structures in the solution $\vec{x}$ can contain physical information. This is particularly true, if the measured distribution has large slopes and second derivatives, which is the case for double differential cross sections. Therefore, one has to check whether the choice τ_L sufficiently preserves all physical information. We will come back to this question in chapter 10.

4.4.2 Algebraic Method

An algebraic approach to unfolding has been proposed by V. BLOBEL [Blo02], from now on referred to as *algebraic method.* At the core of this approach a three step linear transformation $\boldsymbol{T} = \boldsymbol{T}_3 \boldsymbol{T}_2 \boldsymbol{T}_1$ on the unregularized result $\vec{x}$ is introduced, which removes all stastistical correlations between the data points. Moreover, smooth contributions to $\vec{x}$ and such of high curvature will be separated. The entries $(\boldsymbol{T}\vec{x})_i$ of the transformed vector $\boldsymbol{T}\vec{x}$ can then be studied independently, particularly in terms of their statistical significance.

The first transformation $\boldsymbol{T}_1$ is an orthogonal projection onto the eigenvectors of the inverse covariance matrix $\boldsymbol{COV}_{\vec{x}_0}^{-1}$ of the unregularized result $\vec{x}_0$. Since $\boldsymbol{COV}_{\vec{x}_0}^{-1}$ is symmetric, we can find an orthogonal $n \times n$-matrix $\boldsymbol{U}$ and a diagonal $n \times n$-matrix $\boldsymbol{D}$ such, that $\boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{U}^T = \boldsymbol{COV}_{\vec{x}_0}^{-1}$. The trace elements of $\boldsymbol{D}$ are the eigenvalues of $\boldsymbol{COV}_{\vec{x}_0}^{-1}$ and the column vectors of $\boldsymbol{U}$ are the corresponding unit eigenvectors. We now define:

$$\boldsymbol{T}_1 : \mathbb{R}^n \to \mathbb{R}^n, \vec{x} \to \boldsymbol{T}_1 \vec{x} = \boldsymbol{U}^T \vec{x} \tag{4.4.3}$$

The covariance matrix $\boldsymbol{COV}_{\boldsymbol{T}_1 \vec{x}_0}$ of the transformed unregularized result is then diagonal:

$$\boldsymbol{COV}_{\boldsymbol{T}_1 \vec{x}_0} = \boldsymbol{U}^T \cdot \boldsymbol{COV}_{\vec{x}_0} \cdot \boldsymbol{U} \equiv \frac{1}{\boldsymbol{D}} \tag{4.4.4}$$

This means, the coefficients $(\boldsymbol{T}_1 \vec{x}_0)_i$ are uncorrelated; their variance is given by the i-th inverse eigenvalue D_{ii} of $\boldsymbol{COV}_{\vec{x}_0}^{-1}$.

The second transformation $\boldsymbol{T}_2$ scales the components of $\boldsymbol{T}_1 \vec{x}$ such, that their variances become all equal to 1. We define:

$$\boldsymbol{T}_2 : \mathbb{R}^n \to \mathbb{R}^n, \vec{x} \to \boldsymbol{T}_2 \vec{x} = \sqrt{\boldsymbol{D}} \vec{x} \tag{4.4.5}$$

Note, that since $\boldsymbol{COV}_{\vec{x}_0}^{-1}$ is positive semidefinite, its eigenvalues are positive or zero and the square root

$$\left(\sqrt{\boldsymbol{D}}\right)_{lm} = \sqrt{D_{lm}} \tag{4.4.6}$$

is well defined. Then, the covariance matrix $\boldsymbol{COV}_{\boldsymbol{T}_2 \boldsymbol{T}_1 \vec{x}_0}$ is given by:

$$\boldsymbol{COV}_{\boldsymbol{T}_2 \boldsymbol{T}_1 \vec{x}_0} = \sqrt{\boldsymbol{D}} \cdot \frac{1}{\boldsymbol{D}} \cdot \sqrt{\boldsymbol{D}}^T \equiv \boldsymbol{I} \tag{4.4.7}$$

Here, $\boldsymbol{I}$ denotes the unit matrix. All coefficients $(\boldsymbol{T}_2\boldsymbol{T}_1\vec{x}_0)_i$ are uncorrelated and have unit variance.

The third transformation $\boldsymbol{T}_3$ is an orthogonal projection onto the eigenvectors of the transformed curvature matrix $\tilde{\boldsymbol{L}}^2$:

$$\tilde{\boldsymbol{L}}^2 \equiv \left((\boldsymbol{T}_2\boldsymbol{T}_1)^{-1}\right)^T \cdot \boldsymbol{L}^2 \cdot (\boldsymbol{T}_2\boldsymbol{T}_1)^{-1} = \frac{1}{\sqrt{\boldsymbol{D}}} \cdot \boldsymbol{U}^T \cdot \boldsymbol{L}^2 \cdot \boldsymbol{U} \cdot \frac{1}{\sqrt{\boldsymbol{D}}} \qquad (4.4.8)$$

Obviously, $\tilde{\boldsymbol{L}}^2$ is symmetric and can again be diagonalized by an orthogonal matrix $\boldsymbol{B}$ and a diagonal matrix $\boldsymbol{S}$, such that $\boldsymbol{B}\cdot\boldsymbol{S}\cdot\boldsymbol{B}^T = \tilde{\boldsymbol{L}}^2$. Then, $\boldsymbol{S}$ contains the eigenvectors of $\tilde{\boldsymbol{L}}^2$ and the columns of $\boldsymbol{B}$ the corresponding unit eigenvectors. Without loss of generality we can choose $\boldsymbol{S}$ such, that the following *ordering condition* is fulfilled:

$$S_{11} \leq S_{22} \leq ... \leq S_{nn} \qquad (4.4.9)$$

For the third transformation $\boldsymbol{T}_3$ we define:

$$\boldsymbol{T}_3 : \mathbb{R}^n \to \mathbb{R}^n, \vec{x} \to \boldsymbol{T}_3\vec{x} = \boldsymbol{B}^T\vec{x} \qquad (4.4.10)$$

The covariance matrix $\boldsymbol{COV}_{\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1\vec{x}_0}$ of the transformed unregularized result is still equal to the unit matrix:

$$\boldsymbol{COV}_{\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1\vec{x}_0} = \boldsymbol{B}^T \cdot \boldsymbol{I} \cdot \boldsymbol{B} \equiv \boldsymbol{I} \qquad (4.4.11)$$

In order to further study the full transformation

$$\boldsymbol{T} \equiv \boldsymbol{T}_3 \cdot \boldsymbol{T}_2 \cdot \boldsymbol{T}_1 \qquad (4.4.12)$$

we calculate the vectors:

$$\vec{a}_k := \boldsymbol{T}^{-1}\vec{e}_k \forall k < n \qquad (4.4.13)$$

Here, $\vec{e}_k$ denotes the k-th unit vector. We call the vectors $\vec{a}_k$ *modes,* since every solution $\vec{x}$ can now be viewed as linear combination of these $\vec{a}_k$ with statistically independent coefficients:

$$\vec{x} = \sum_k (\boldsymbol{T}\vec{x})_k \, \vec{a}_k \qquad (4.4.14)$$

We now study the coefficient $(\boldsymbol{T}\vec{x}_0)_k$ of the unregularized result for every mode independently. In particular, the coefficients can be classified according to their statistical significance. This is done by comparing the absolute mean value $|\boldsymbol{T}\vec{x}_0|$ of each coefficient to its variance, which is 1. We only treat those modes as statistically significant, for which the following *significance criterion* is fulfilled:

$$|\boldsymbol{T}\vec{x}_0| > \zeta \simeq 3.84 \qquad (4.4.15)$$

Obviously, the value of the parameter ζ can be used for steering. For this analysis, the rather high value of $\zeta \simeq 3.84$ has been chosen; this corresponds to the width of the 95% confidence interval of a normally distributed variable with known variance 1. The number of statistically significant modes is called the *effective dimension n_0,* which is always less than or equal to n.

For an example, see figure 10.2.3 in chapter 10 which will be explained in all detail there. Essentially, it shows the measured and the expected double-differential cross

section for neutral current *ep*-scattering, both transformed according to (4.4.12). The descent of the coeffientents $(\boldsymbol{T}\vec{x}_0)_k$ with k is clearly visible as well as dropping below the significance threshold $\zeta \simeq 3.84$.

Note, that the study of the significance of individual modes - and in particular the effective dimension n_0 - depends on the curvature measure $\boldsymbol{L}^2$, which enters the definition of the third transformation $\boldsymbol{T}_3$. Consider, that $\boldsymbol{T}_3$ is orthogonal for every choice of $\boldsymbol{L}^2$ and will thus always give a unit covariance matrix $\boldsymbol{COV}_{\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1\vec{x}_0} = \boldsymbol{I}$, while $\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1\vec{x}_0$ can vary significantly with $\boldsymbol{L}^2$. However, if the curvature measure $\boldsymbol{L}^2$ is well chosen, an insignificant coefficient $(\boldsymbol{T}\vec{x}_0)_i$ will belong to a mode $\vec{a}_i$ of high curvature, meaning that it corresponds to a large eigenvalue S_{ii} of $\tilde{\boldsymbol{L}}^2$.

The effect of $\boldsymbol{T}$ on a regularized solution $\vec{x}(\tau)$ can be calculated from (4.2.15). Let $\hat{\boldsymbol{L}}^2$ denote the fully transformed curvature matrix:

$$\hat{\boldsymbol{L}}^2 \equiv \left((\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1)^{-1}\right)^T \cdot \boldsymbol{L}^2 \cdot (\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1)^{-1} \tag{4.4.16}$$

We obtain by direct computation:

$$\boldsymbol{T}\vec{x}_\tau = \boldsymbol{B}^T \cdot \sqrt{\boldsymbol{D}} \cdot \boldsymbol{U}^T \cdot \left[\frac{\boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \vec{N}^{data}}{\boldsymbol{E}_0^{-1} + \tau^2 \boldsymbol{L}^2} + \tau^2 \boldsymbol{E}_0 \boldsymbol{L}^2 \vec{x}^{Theory}\right]$$

$$\boldsymbol{T}\vec{x}_\tau = \boldsymbol{B}^T \cdot \sqrt{\boldsymbol{D}} \cdot \boldsymbol{U}^T \cdot \frac{\boldsymbol{E}_0^{-1} \cdot \boldsymbol{U} \cdot \sqrt{\boldsymbol{D}}^{-1} \cdot \boldsymbol{B}}{\boldsymbol{E}_0^{-1} + \tau^2 \boldsymbol{L}^2} (\boldsymbol{T}\vec{x}_0) + \tau^2 \hat{\boldsymbol{L}}^2 \left(\boldsymbol{T}\vec{x}^{Theory}\right)$$

$$\boldsymbol{T}\vec{x}_\tau = \frac{1}{\boldsymbol{I} + \tau^2 \cdot \boldsymbol{S}} (\boldsymbol{T}\vec{x}_0) + \tau^2 \hat{\boldsymbol{L}}^2 \left(\boldsymbol{T}\vec{x}^{Theory}\right) \tag{4.4.17}$$

For now, set $\vec{x}^{Theory} = 0$. Then, the regularization can be viewed as a damping of the coefficients $(\boldsymbol{T}\vec{x}_0)_i$ with high curvature eigenvalues S_{ii}:

$$(\boldsymbol{T}\vec{x}_\tau)_i = \frac{1}{1 + \tau^2 \cdot S_{ii}} (\boldsymbol{T}\vec{x}_0)_i \tag{4.4.18}$$

This means, that the contribution of modes $\vec{a}_k$ with $\tau^2 \cdot S_{kk} \gtrsim 1$ will effectively be cut away. Therefore, the effective dimension n_0 suggests a choice τ_A for the regularization parameter. By summing up the damping factors $\frac{1}{1+\tau^2 \cdot S_{ii}}$ we obtain:

$$\eta(\tau) = \sum_i \frac{1}{1 + \tau^2 \cdot S_{ii}} \tag{4.4.19}$$

Now, we set:

$$\eta(\tau_A) = n_0 \tag{4.4.20}$$

This means, that we identify the number of effective dimensions n_0 from a significance analysis of the unregularized modes, then (4.4.19) will be inverted to find τ_A.

4.4.3 Global Correlation Method

A third choice for the regularization parameter τ_C can be motivated by using the notion of global correlation [Blob]. The idea is to find a quantity, that measures the correlation of the event count x_i in a bin $\mathcal{B}_i$ to its whole environment, i.e. to the event

counts x_j in all surrounding bins $\mathcal{B}_j$ with $j \neq i$. Such a correlation measure is then naturally a τ-dependent function. Thus, an optimum choice for τ can be obtained from minimizing this function with respect to τ.

To define such a correlation measure for bin $\mathcal{B}_i$, we scan all linear combinations $\boldsymbol{\lambda}_i = \sum_{j \neq i} \alpha_j \cdot x_j$ of data points ($\alpha_j \in \mathbb{R}$) for the combination $\boldsymbol{\lambda}_i^{max}$ which maximizes the correlation $\mathrm{Corr}\,(x_i, \boldsymbol{\lambda}_i)$ with bin $\mathcal{B}_i$. More formally, let the generator level bin grid $\mathcal{G}_{gen}$ have n bins. For every bin $\mathcal{B}_i$, we define the set $\boldsymbol{\Lambda}_i$ of linear combinations of all environmental data points x_j with $j \neq i$:

$$\boldsymbol{\Lambda}_i := \left\{ \boldsymbol{\lambda}_i \,\middle|\, \boldsymbol{\lambda}_i = \sum_{j \neq i} \alpha_j \cdot x_j,\ \alpha_j \in \mathbb{R} \right\} \tag{4.4.21}$$

Then, the *global correlation* ρ_i in bin $\mathcal{B}_i$ is defined as the maximum of correlations between x_i and any linear combination $\boldsymbol{\lambda}_i \in \boldsymbol{\Lambda}_i$ [Jam, Bloc]:

$$\rho_i := \max_{\boldsymbol{\lambda}_i \in \boldsymbol{\Lambda}_i} \left(\mathrm{Corr}\,(x_i, \boldsymbol{\lambda}_i) \right) \tag{4.4.22}$$

While the definition (4.4.22) seems to be involved, the global correlation can easily be computed from the covariance matrix $\boldsymbol{COV}_{\vec{x}}$ of $\vec{x}$ [Bloc]:

$$\rho_i = \sqrt{1 - \frac{1}{(\boldsymbol{COV}_{\vec{x}})_{ii} \cdot \left(\boldsymbol{COV}_{\vec{x}}^{-1} \right)_{ii}}} \tag{4.4.23}$$

Note, that by construction ρ_i is always a real number between 0 and 1.

The notion of global correlation is useful, since it provides information on how a single data point is correlated to its whole environment. Moreover, it is sensitive to the choice of τ. For very small τ (*under-smoothing*), the global correlation $\rho_i(\tau)$ will show rather large values (close to 1), which will decrease with rising τ. This is due to the large negative bin-to-bin-correlations which are introduced by migration. For very large τ (*over-smoothing*), the behaviour of $\rho_i(\tau)$ depends on the curvature measure. For size regularization, $\rho_i(\tau)$ tends to 0, while for derivational and curvature regularization $\rho_i(\tau)$ rises again with τ. This is, because these curvature measures mainly constrain the first and second derivative of the measured distribution, introducing positive bin-to-bin correlations.

Care needs to be taken, if a bin $\mathcal{B}_i$ is situated close to a bin $\mathcal{B}_j$ of very low efficiency (*low efficiency bin*). A low efficiency bin is poorly constrained by data. Therefore, due to the smoothing prescription, its value x_j can be fully determined by its neighbouring bin x_i. In this case, a large positive correlation ρ_{ij} would arise. These correlations would strongly contribute to the global correlation of bin $\mathcal{B}_i$.

We can avoid this problem by noting, that many bins of low efficiency have to be excluded from quotation anyhow, since they suffer from large systematic uncertainties. Therefore, we define the *reduced global correlation* which essentially ignores all the bins that lack our interest. Pratically, only those bins should enter our notion of reduced global correlation, that will be quoted in the end. Starting from $\boldsymbol{COV}_{\vec{x}}$ we define the submatrix $\boldsymbol{COV}_{\vec{x}}^{red}$ by removing the columns and rows of those bins that are not to be quoted. Then we set:

$$\rho_i^{red} := \sqrt{1 - \frac{1}{\left(\boldsymbol{COV}_{\vec{x}}^{red} \right)_{ii} \cdot \left(\left(\boldsymbol{COV}_{\vec{x}}^{red} \right)^{-1} \right)_{ii}}} \tag{4.4.24}$$

This quantity shows the same properties as ρ_i does, besides being more reliable at the edges of the bin grid. On the other hand, it is not defined for non-quotable bins, particularly overflow bins. For brevity, we will use the term "global correlation" and its symbol ρ_i to refer to the definition given in (4.4.24).

In the context of derivational and curvature regularization, this behaviour of $\rho_i(\tau)$ can be used to define an optimum regularization parameter τ_C by minimizing $\rho_i(\tau)$ as a function of τ. To achieve such a minimization for the bin grid as a whole, we introduce some averaging notions. We define the *maximum global correlation* as:

$$\rho_{max}(\tau) = \max_i \left(\rho_i^{red}(\tau) \right) \tag{4.4.25}$$

We define the *averaged global correlation* as:

$$\rho_{av}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \left(\rho_i^{red}(\tau) \right) \tag{4.4.26}$$

We define the *averaged squared global correlation* as:

$$\rho_{avsq}(\tau) = \frac{1}{n} \sum_{i=1}^{n} \left(\rho_i^{red}(\tau) \right)^2 \tag{4.4.27}$$

Note, that the sum in these definitions should run only over those bins, for which the reduced global correlation is defined.

By minimizing each of these functions, similar values τ_C for the regularization parameter can be obtained. For an example, see figure 10.2.7 in chapter 10. For a two dimensional unfolding analysis performed there, it shows the averaged global correlation $\rho_{av}(\tau)$ and the averaged squared global correlation $\rho_{avsq}(\tau)$ as a function of the regularization parameter τ. The minimum τ_C is clearly visible.

The choice τ_C for the regularization level minimizes bin-to-bin correlations. This means, that the covariance matrix $\boldsymbol{COV}_{\vec{x}}$ will typically be rather diagonal. This is useful, whenever the researcher's interest lies mainly in the shape of the measured distribution rather than statistical uncertainties and correlations.

4.4.4 Comparison of Regularization Levels

The three different methods described in the preceding subsections each provides their own regularization parameter, τ_L, τ_A and τ_C. They each have advantages and drawbacks.

The algebraic method aims at finding a solution $\vec{x}$, that maximizes the information captured from the data while leveling out unphysical fluctuations. This is reflected in cutting away only those entries from $\boldsymbol{T}\vec{x}_0$ that are truly insignificant. Therefore, the choice of τ_A is rather conservative. Some curvature might still remain in the solution $\vec{x}$, as long as it is statistically significant. Moreover, data points x_i might still be negatively correlated. While the statistical uncertainties will typically be rather large, they are inconclusive. This is, because they are contaminated by contributions from bin to bin correlations.

The result τ_C from the global correlation is typically higher than τ_A. It minimizes the bin to bin correlations, potentially sacrificing some statistically significant information. We end up with a solution $\vec{x}$, that is smoother than that from the algebraic

method. Moreover, it shows smaller and more realistic statistical uncertainties. It is the method of choice, if one needs to visualize the shape of the distribution $\vec{x}$ or is interested in the statistical accuracy of the measurement.

The L curve method tries to find a trade-off between data compatability (χ_A^2) and curvature (χ_L^2). Its output τ_L is therefore typically close to the result of the global correlation method.

4.5 The Choice of the Bin Grid

The notions of purity p_i and stability s_i reflect both the detector performance and the characteristics of the chosen bin grid. Therefore, a careful choice of the involved bin sizes is imperative. For the generator level bin grid $\mathcal{G}_{gen}$, a number of conditions can be listed, which should be fulfilled for a sucessful unfolding procedure.

4.5.1 Detector Resolution

At first, the bin grid $\mathcal{G}_{gen}$ should reflect the resolution scale of the detector. In principle, unfolding provides a technology to de-blur measured data to scales *far below* the detector resolution. However, the success of such a procedure depends strongly on the reliability of the detector simulation. The smaller the bin size, the stronger is the influence of the response matrix $\boldsymbol{N}^{MC}$ on the unfolding result $\vec{x}$, and the stronger the systematic uncertainty on the entries N_{ji}^{MC} will affect $\vec{x}$.

From an experimental viewpoint, this is clear, because the information contained in the measured data $\vec{N}^{data}$ is limited, regardless of the chosen bin size. Mathematically, it is reflected in formula (4.8.16): If the bin sizes are smaller than the detector resolution, large off-diagonal elements E_{ji} of $\boldsymbol{E} = \left(\boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \boldsymbol{A} \right)^{-1}$ will emerge. These will increase the dependence of the unfolding result $\vec{x}$ on the elements N_{ji}^{MC}, see (4.8.16). Therefore, by error propagation, large systematic uncertainties will be introduced on the result $\vec{x}$.

A choice for the bin size can be made exploiting the notion of stability. For one dimensional problems, one can choose a bin grid $\mathcal{G}_{gen}$ such, that the stability roughly takes on a constant value s_{1D} for every bin $\mathcal{B}_i$:

$$s_i \simeq s_{1D} \tag{4.5.1}$$

To motivate a reasonable choice for s_{1D}, consider the simple case, in which a measured quantity η^{meas} is GAUSSIAN distributed around its true value $\mu = \eta^{true}$ with a variance $\sigma^2 = \Lambda^2$. Here, Λ reflects the detector resolution scale, which in this example is assumed to be constant throughout the bin grid. Then, the stability of $s_{1D} \simeq 80\%$ reflects the reasonable choice of $\sim 2.5\Lambda$ for the bin width:

$$s_{1D} \simeq 80\% \tag{4.5.2}$$

Note, that the notion of stability is applicable even to those frequent cases, in which detector effects are neither GAUSSIAN distributed nor show a constant resolution scale.

For the two dimensional case, the stability should be both constant throughout the bin grid $\mathcal{G}_{gen}$ and independent of the direction:

$$s_{i,j} \simeq s_{2D} \forall i,j \tag{4.5.3}$$

$$s_i^k \simeq s_j^l \simeq s_{1D} \forall i,j \ \forall k,l \in \{1,2\} \tag{4.5.4}$$

Note, that these two conditions correspond to the condition of regional and directional conformity, (4.3.14) and (4.3.15).

To motivate a choice for s_{2D}, assume two quantities η_1^{meas} and η_2^{meas} which are GAUSSIAN distributed around their true values $\mu_1 = \eta_1^{true}$ and $\mu_2 = \eta_2^{true}$ and with variances $\sigma_1^2 = \Lambda_1^2$ and $\sigma_2^2 = \Lambda_2^2$ reflecting two constant resolution scales. The choice of rectangular bins with the widths $\sim 2.5\Lambda_1$ and $\sim 2.5\Lambda_2$ then leads to a stability of $s_{2D} \simeq 60\%$. Therefore, for the two dimensional case we can reasonably choose:

$$s_{2D} \simeq 60\% \tag{4.5.5}$$

Note, that for underflow and overflow bins the conditions (4.5.1), (4.5.5) and (4.5.4) are typically not fulfilled.

4.5.2   Effective Dimension

The second constraint on the bin grid concerns the *effective dimension n_0*. In subsection 4.4.2 it has been pointed out, that n_0 counts the statistically independent data points of the measurement. Therefore, n_0 can be used to estimate the optimum number of bins n in the bin grid.

Note, however, that the effective dimension n_0 is a direct property of the transformed result $\boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1\vec{x}$ and such depends on the response matrix $\boldsymbol{A}$, the curvature measure, the data distribution and its statistical uncertanties. This shows, that the optimum binning is not a mere question of the detector resolution. It also depends on the functional shape under investigation ($\vec{N}^{data}$), the available data statistics ($\boldsymbol{COV}_{\vec{N}^{data}}$) and the researcher's notion of curvature ($\vec{x} \to \mathcal{C}(\vec{x})$).

Note further, that the effective dimension n_0 depends on the number of bins n chosen in the first place, and that we have $n_0(n) < n$. This reflects the natural correlations between the bins. These can never be completely suppressed, regardless of the bin size and the bins are never truly statistically independent (i.e. $n_0(n) = n$). Therefore, to adjust the number of bins to $n_0(n)$, i.e. $n \to n' \simeq n_0(n)$ contains the risk of loosing information, since we will have $n_0(n') < n_0(n)$. To avoid the dependence on n and loss of information, n should be chosen such that we have:

$$n \gtrsim 1.5 \cdot n_0(n) \tag{4.5.6}$$

In some cases bin grids contain n_{ZE} bins with vanishing efficieny ("zero efficiency bins"), see section 4.5.3. This is often the case in two dimensional scenarios, if the rectangular bin grid reaches out of the acceptance region of the detector. The measured data do not contain information on these n_{ZE} bins. As described in section 4.5.3, these bins must be removed prior to unfolding. In this case, the number n in this subsection must be replaced with:

$$n = n_{all} - n_{ZE} \tag{4.5.7}$$

4.5.3   Quotable and Non-Quotable Bins

Once a generator level bin grid $\mathcal{G}_{gen}$ is found, not all bins are alike. In particular, there are bins $\mathcal{B}_i$ for which the unfolding result x_i must not be quoted, since we cannot trust it. We refer to these as *non-quotable (NQ) bins*.

The *overflow (OF) bins* will be regarded as non-quotable. This is, because every overflow bin covers an infinite phase space region, which the detector model must naturally fail to simulate. Of course, this does not relax the necessity to carry all overflow bins through the unfolding process.

In many cases the efficiency ϵ varies strongly throughout the bin grid. Often, this is due to the boundaries of the detector acceptance, which lead to "blind spots" in some phase space regions of the bin grid. One can define an *efficiency threshold ϵ_{tr}* and exclude all those bins from quotation that fail to satisfy the following *efficiency condition:*

$$\epsilon_i \geq \epsilon_{tr} \tag{4.5.8}$$

We call such bins *low efficiency (LE) bins*, as long as they are not overflow bins. We cannot trust the unfolding result x_i in a low efficiency bin $\mathcal{B}_i$, since it is afflicted with a rather large systematic uncertainty from the efficiency modelling of the Monte Carlo simulation. Of course, the threshold ϵ_{tr} is a tunable parameter, the choice of which depends on the problem under investigation. The question, whether all low efficiency bins have to be carried trough the unfolding process, will be addressed in the section 4.5.5.

A more intricate problem emerges from those low efficiency bins, for which the efficiency exactly vanishes. These are called *zero efficiency (ZE) bins.* The solution x_i of a zero efficiency bin $\mathcal{B}_i$ is not constrained by any data and thus $\mathcal{B}_i$ has to be excluded from unfolding. The procedure for that is described in section 4.5.5.

For further clarity, we introduce the following notation: Let n_{all} denote all generator level bins, n_{OF} the number of overflow bins, n_{LE} the number of low efficiency bins, n_{ZE} the number of zero efficiency bins, n_{QO} the number of quotable bins, n_{NQ} the number of non-quotable bins and n the number of generator level bins entering the unfolding process. Then we have:

$$n = n_{all} - n_{ZE} \tag{4.5.9}$$

$$n_{QO} = n_{all} - n_{NQ} = n_{all} - n_{OF} - n_{LE} \tag{4.5.10}$$

4.5.4   The Detector Level Bin Grid

In general, the choice of the detector level bin grid $\mathcal{G}_{det}$ is less intricate. A general rule is, to choose $\mathcal{G}_{det}$ finer than $\mathcal{G}_{gen}$, in order to avoid loss of resolution [Blo02]. Typically one uses for every dimension twice as many bins on the detector level than one uses on the generator level. For two dimensions, this leaves us with four detector level bins for each non-overflow generator level bin.

Attention needs to be paid to those detector level bins, which do not contain any data, i.e. for which $N_i^{data} = 0$. We call them *empty (EM) bins*. Since the event counts N_i^{data} are assumed to be POISSON-distributed, these bins have vanishing statistical uncertainty; this leads to a covariance matrix $\boldsymbol{COV}_{\vec{N}^{data}}$ that can not be inverted. The empty bins have to be excluded from unfolding.

For further clarity, we introduce the following notation: Let m_{all} denote all detector level bins, m_{OF} the number of overflow bins on detector level, m_{EM} the number of empty bins and m the number of detector level bins that enter the unfolding process. Then we have:

$$m = m_{all} - m_{EM} \tag{4.5.11}$$

4.5.5   Treatment of Low Efficiency Bins

While it is clearly imperative to carefully account for every overflow bin, the question arises, whether every low efficiency bin needs to be carried trough the whole unfolding process. Two possible approaches are available: To cut out the low efficiency bins *prior* to unfolding as it is done for the zero efficiency bins ("cut-first-approach"), or to discard the low efficiency bins after the unfolding process ("cut-last-approach").

The advantage of the first method is, that only those bins enter the unfolding process which are of physical interest. The behaviour of the low efficiency bins during the regularization process if often difficult to understand. However, the bin grid will no longer be rectangular, which makes the definition of the curvature measure slightly more complex. Formula (4.5.9) becomes:

$$n = n_{all} - n_{LE} \tag{4.5.12}$$

4.5.6   Internal Mappings

In order to cut out generator level bins prior to unfolding, an internal bin mapping needs to be found[2]:

$$\nu : \{1, ..., n_{all}\} \to \{1, ..., n\}, \ k \to \nu(k) \tag{4.5.13}$$

We find a map ν such, that the index k of a low efficiency bin $\mathcal{B}_k$ is mapped onto the bin number $\nu(k)$ of the closest overflow bin $\mathcal{B}_{\nu(k)}$. In most cases, this choice is natural and unique. If $\mathcal{B}_k$ is not a low efficiency bin, then k is mapped onto itself. Then, the unfolding matrix $\boldsymbol{N}^{MC}$ has to be recalculated, which essentially means adding the column of bin $\mathcal{B}_k$ to the column of the overflow bin $\mathcal{B}_{\nu(k)}$ and removing it thereafter. Moreover, N_k^{Lost} needs to be added to $N_{\nu(k)}^{Lost}$, then has to be removed.

In order to cut out detector level bins prior to unfolding, an internal bin mapping needs to be found:

$$\mu : \{1, ..., m_{all}\} \to \{1, ..., m\}, \ k \to \mu(k) \tag{4.5.14}$$

We find a map μ such, that the index k of an empty bin $\mathcal{B}_k$ is mapped onto the bin number $\mu(k)$ of the closest overflow bin $\mathcal{B}_{\mu(k)}$. In most cases, this choice is natural and unique. If $\mathcal{B}_k$ is not an empty bin, then k is mapped onto itself. Then, the unfolding matrix $\boldsymbol{N}^{MC}$ has to be recalculated, which essentially means adding the row of bin $\mathcal{B}_k$ to the row of that overflow bin $\mathcal{B}_{\mu(k)}$ and removing it thereafter. Moreover, N_k^{data} needs to be added to $N_{\mu(k)}^{data}$, then has to be removed. The curvature matrix $\boldsymbol{L}^2$ needs to be recalculated, too.

4.6   Treatment of Background Sources

The general approach to treat background in this analysis is to subtract it from the data distribution *before* unfolding. Let $\vec{N}^{raw}$ be the measured data distribution and

---

[2]The mapping ν is not to be confused with the index mapping ν_{2D} from (4.3.6) and (4.3.7).

$\vec{N}^{bgr}$ the reconstructed distribution of some background simulation. Then, the data distribution $\vec{N}^{data}$ which enters the unfolding procedure is given by:

$$\vec{N}^{data} = \vec{N}^{raw} - \frac{\mathcal{L}_{data}}{\mathcal{L}_{bgr}} \cdot \vec{N}^{bgr} \equiv \vec{N}^{raw} - f_{bsc} \cdot \vec{N}^{bgr} \qquad (4.6.1)$$

Here, the symbols $\mathcal{L}_{data}$ and $\mathcal{L}_{bgr}$ denote the luminosities of the data sample and the background simulation, respectively. f_{bsc} is called the *background scale factor.*

4.7   Bin-by-Bin Unfolding

The traditional way to treat detector effects is a mere bin wise efficiency correction. For this, the same bin grid needs to be used on generator and detector level, i.e. $\mathcal{G}_{gen} = \mathcal{G}_{det}$. The bin-by-bin efficiency ϵ_i^{bbb} is determined for every bin, and every data point x_i is obtained by the following procedure:

$$x_i := \frac{1}{\epsilon_i^{bbb}} \cdot N_i^{data} \qquad (4.7.1)$$

Here, N_i^{data} refers to the i-th data point of the background free measurement, see (4.6.1). This procedure is often regarded as *bin-by-bin unfolding*, since it can be viewed as an unfolding procedure with a diagonal response matrix $\boldsymbol{A}^{bbb}$:

$$\left(\boldsymbol{A}^{bbb}\right)_{ji} = \delta_{ji} \cdot \epsilon_i^{bbb} \qquad (4.7.2)$$

4.7.1   Draw Backs of Bin-by-Bin Unfolding

Bin-by-bin unfolding not only compensates efficiency losses, it even accounts to some extend for migration. Under the assumption of a reliable Monte Carlo simulation, the migration effects will also show up in the reconstructed model distribution $\vec{N}^{meas}$ and will thus be described by the bin-by-bin efficiency ϵ_i^{bbb}. This is exactly, because the bin-by-bin efficiency is *not* a true efficiency. Instead we have:

$$\epsilon_i^{bbb} = \epsilon_i + \frac{N_i^{Smearin} - N_i^{Smearout}}{N_i^{gen}} \qquad (4.7.3)$$

Here, $N_i^{Smearin}$ denotes the number events generated in some bin $j \neq i$ and reconstructed in bin i, while $N_i^{Smearout}$ denotes the events that have been generated in bin i and reconstructed in some bin $j \neq i$. This means among other things, that the quantity ϵ_i^{bbb} can have values grater than 1, and that ϵ_i^{bbb} does contain information about migration.

We see, that the inadequacy of the bin-by-bin method does not lie in the insensitivity to migration effects, it rather lies in the wrong treatment of correlations. Since $\boldsymbol{A}^{bbb}$ is diagonal, the covariance matrix $\boldsymbol{COV}_{\vec{x}}$ will also be diagonal, which contradicts the nature of migration.

A second deficiency of bin-by-bin method is, that the statistical uncertainties cannot be trusted a priori. This is the case, whenever there is a significant imbalance

between immigration and emigration, i.e. if $\left| N_i^{Smearin} - N_i^{Smearout} \right| \gg 0$. Consider, for example, a distribution $\vec{N}^{true}$ that steeply falls with its entry index i:

$$N_1^{true} > N_2^{true} > ...$$

This is true for many differential cross section measurements. Then, typically $N_i^{Smearin} - N_i^{Smearout} > 0$. Further, assume the overall efficiency to be close to one ($\epsilon_i \simeq 1$) and well known ($\delta(\epsilon_i) \simeq 0$). Then, the true relative statistical uncertainty $\sigma\left(N_i^{true}\right)$ will be underestimated by the one obtained from bin-by-bin unfolding:

$$\delta\left(x_i^{bbb}\right) = \delta\left(N_i^{rec}\right) = \frac{1}{\sqrt{N_i^{rec}}} =$$

$$\frac{1}{\sqrt{N_i^{true} + N_i^{Smearin} - N_i^{Smearout}}} < \frac{1}{\sqrt{N_i^{true}}} = \delta\left(N_i^{true}\right)$$

For distributions $\vec{N}^{true}$ that show local maxima (*peaks*), statistical uncertainties can be overestimated for a similar reason.

4.7.2 Global Correlation Method as Cross Check for Bin-By-Bin Unfolding

The bin-by-bin method can be viewed as an unfolding technique with *implicit regularization*. Recall, that the global correlation method described in subsection 4.4.3 provides a measure of regularization τ_C, that minimizes the correlations of $\vec{x}$ in a global manner. This means, that the off-diagonal elemens in $\boldsymbol{COV}_{\vec{x}}$ will be minimized. While the global correlation method achieves the diagonalization of $\boldsymbol{COV}_{\vec{x}}$ by a smoothing procedure, the bin-by-bin method sets the off-diagonal elements of $\boldsymbol{A}^{bbb}$ to zero in the first place, leading to a diagonal $\boldsymbol{COV}_{\vec{x}}$.

We see, that regardless of the deficiencies of the bin-by-bin method, it implicitly introduces a regularization. The level of this regularization cannot be expressed in terms of a precise value of τ. However, the result $\vec{x}^{bbb}$ of the bin-by-bin method can meaningfully be compared to the result $\vec{x}_{\tau_C}$ of a full unfolding procedure using τ_C according to the global correlation method.

This provides a valuable cross check for the bin-by-bin method and its ability to quantify the statistical uncertainties correctly.

4.8 Treatment of Systematic Uncertainties

The unfolding procedure provides a tool to directly quantify the statistical errors $\Delta^{stat}x_i$ of the unfolded data points x_i as well as their bin-to-bin correlations $\rho_{ij} := \boldsymbol{Corr}(x_i, x_j)$, see (4.8.2). Systematic uncertainties $\Delta^{sys}x_i$ can be determined by error propagation [Scha]. This is implemented by S. SCHMITT in the ROOT-package TUnfoldSys [Schb].

There are many types of systematic uncertainties. The unfolding matrix $\boldsymbol{A}$ gives two types of systematic uncertainties, *uncorrelated model uncertainties* $\Delta^{unc}x_i$ and *correlated model uncertainties* $\Delta^{\lambda}x_i$ for multiple error sources λ. The background distribution $\vec{N}^{bgr}$ introduces the *background uncertainty* $\Delta^{bgr}x_i$ and the correlated *background scale uncertainty* $\Delta^{bsc}x_i$.

4.8.1   Treatment of Data and Background Uncertainties

In this analysis, we adopt the viewpoint of a *bias-free measurement*. That means, the *raw measurement* $\vec{N}^{raw}$ only suffers from statistical uncertainties and is free from any systematic deviations. This perspective can be motivated by noting, that the data distribution is compared to a reconstructed Monte Carlo simulation $\vec{N}^{rec}$ and any flaw in the measurement process will ideally be reproduced by the simulation. Therefore, rather the simulation than the data will be afflicted with possible systematic errors. We will postpone their treatment to the next subsection.

The event counts N_i^{raw} of the raw measurement are uncorrelated and POISSON-distributed, in contrast to the data points x_i of the unfolded solution $\vec{x}$. Therefore, we have $\Delta N_i^{raw} \simeq \sqrt{N_i^{raw}}$ and its covariance matrix $\boldsymbol{COV}_{\vec{N}^{raw}}$ is diagonal.

Concerning the background, there are mainly two error sources to be considered, its statistical error ΔN_i^{bgr} in each bin and its scale uncertainty. Any uncertainty on the background ΔN_i^{bgr} or Δf_{bsc} will directly propagate to $\vec{N}^{data}$. Therefore, the formula (4.2.16) stays valid even in the presence of background sources. Note, however, that the background scale f_{bsc} introduces correlations between the data points N_i^{data}, so that $\boldsymbol{COV}_{\vec{N}^{data}}$ will no longer be exactly diagonal.

As for the data, the background event counts N_i^{bgr} are uncorrelated and POISSON-distributed, $\Delta N_i^{bgr} \simeq \sqrt{N_i^{bgr}}$, its covariance matrix $\boldsymbol{COV}_{\vec{N}^{bgr}}$ is diagonal. To compute $\boldsymbol{COV}_{\vec{N}^{data}}$, we use error propagation with $\frac{\partial}{\partial f_{bsc}} N_i^{data} = N_i^{bgr}$ and obtain:

$$\boldsymbol{COV}_{\vec{N}^{data}} = \boldsymbol{COV}_{\vec{N}^{raw}} + f_{bsc}^2 \cdot \boldsymbol{COV}_{\vec{N}^{bgr}} + \Delta f_{bsc}^2 \cdot \left(\left(\vec{N}^{bgr} \right) \cdot \left(\vec{N}^{bgr} \right)^T \right) \quad (4.8.1)$$

Each of the terms in (4.8.1) can be individually propagated with (4.2.16). This enables us to study their impact on the solution $\vec{x}$ separately. Following (4.2.16), we define the *raw statistical uncertainty* $\Delta^{raw} x_i$:

$$\Delta^{raw} x_i = \sqrt{\left(\left[\boldsymbol{EA}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right] \cdot \boldsymbol{COV}_{\vec{N}^{raw}} \cdot \left[\boldsymbol{EA}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right]^T \right)_{ii}} \quad (4.8.2)$$

The *statistical background uncertainty* $\Delta^{bgr} x_i$ is given by:

$$\Delta^{bgr} x_i = \sqrt{\left(\left[\boldsymbol{EA}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right] \cdot f_{bsc}^2 \boldsymbol{COV}_{\vec{N}^{bgr}} \cdot \left[\boldsymbol{EA}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right]^T \right)_{ii}} \quad (4.8.3)$$

The *statistical scale uncertainty* $\Delta^{bsc} x_i$ is given by:

$$\Delta^{bsc} x_i = \sqrt{\left(\left[\boldsymbol{EA}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right] \cdot \Delta f_{bsc}^2 \left(\left(\vec{N}^{bgr} \right) \left(\vec{N}^{bgr} \right)^T \right) \cdot \left[\boldsymbol{EA}^T \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right]^T \right)_{ii}} \quad (4.8.4)$$

We refer to the quadratic sum of these terms as *statistical uncertainty* $\Delta^{stat} x_i$ and directly obtain with 4.2.16:

$$\Delta^{stat} x_i = \sqrt{\boldsymbol{E}_{ii}} \quad (4.8.5)$$

4.8.2 Treatment of Uncorrelated Model Uncertainties

Since the scaled unfolding matrix $\boldsymbol{N}^{MC}$ from (4.2.3) is determined by a Monte Carlo simulation, each entry N_{jk} comes with an absolute statistical uncertainty $\Delta^{stat}\left(N_{jk}^{MC}\right)$, which is approximately POISSON-distributed:

$$\Delta^{stat}\left(N_{jk}^{MC}\right) \simeq \sqrt{N_{jk}^{MC}} \tag{4.8.6}$$

We treat each of these uncertainties as an uncorrelated systematic error source, which gives a covariance matrix $\boldsymbol{COV}_{\vec{x}}^{jk}$ for every element N_{jk}^{MC} of the matrix $\boldsymbol{N}^{MC}$. By standard error propagation, we obtain:

$$\left(\boldsymbol{COV}_{\vec{x}}^{jk}\right)_{lm} := \frac{\partial x_l}{\partial N_{jk}^{MC}} \cdot \Delta^{stat}N_{jk}^{MC} \cdot \frac{\partial x_m}{\partial N_{jk}^{MC}} \tag{4.8.7}$$

We define the total *uncorrelated uncertainty* simply as the quadratic sum of the contributions of the matrix elements, or, more formally:

$$(\boldsymbol{COV}_{\vec{x}}^{unc})_{lm} := \sum_{jk} \left(\boldsymbol{COV}_{\vec{x}}^{jk}\right)_{lm} \tag{4.8.8}$$

Note, that for the response matrix $\boldsymbol{A}$ we have:

$$A_{ji} = \frac{N_{ji}^{MC}}{N_i^{gen}} = \frac{N_{ji}^{MC}}{\sum_k^{incl} N_{ki}^{MC}} \tag{4.8.9}$$

Here, the sum in the denominator runs over all detector level bins including the overflow bins. From (4.8.9) we obtain:

$$\frac{\partial A_{ab}}{\partial N_{ij}} := \delta_{bj} \cdot \frac{1}{\sum_k^{incl} N_{kb}^{MC}} \cdot (\delta_{ai} - A_{ab}) \tag{4.8.10}$$

Therefore, (4.8.7) gives:

$$\left(\boldsymbol{COV}_{\vec{x}}^{jk}\right)_{lm} = \sum_{abcd} \frac{\partial x_l}{\partial A_{ab}} \cdot \frac{\partial A_{ab}}{\partial N_{jk}^{MC}} \cdot \Delta^{stat}N_{jk}^{MC} \cdot \frac{\partial x_m}{\partial A_{cd}} \cdot \frac{\partial A_{cd}}{\partial N_{jk}^{MC}}.$$

$$\left(\boldsymbol{COV}_{\vec{x}}^{jk}\right)_{lm} = \sum_{ac} \frac{\partial x_l}{\partial A_{ak}} \cdot \frac{\partial A_{ak}}{\partial N_{jk}^{MC}} \cdot \Delta^{stat}N_{jk}^{MC} \cdot \frac{\partial x_m}{\partial A_{ck}} \cdot \frac{\partial A_{ck}}{\partial N_{jk}^{MC}}. \tag{4.8.11}$$

For the *uncorrelated systematic uncertainty* $\Delta^{uncorr}x_i$ we therefore obtain with (4.8.8):

$$\Delta^{uncorr}x_i = \sqrt{\sum_{ijkl} \frac{\partial x_i}{\partial A_{ak}} \cdot \frac{\partial A_{ak}}{\partial N_{jk}^{MC}} \cdot \Delta^{stat}N_{jk}^{MC} \cdot \frac{\partial x_i}{\partial A_{ck}} \cdot \frac{\partial A_{ck}}{\partial N_{jk}^{MC}}} \tag{4.8.12}$$

To calculate the error propagation from $\boldsymbol{A}$ to $\vec{x}$, we observe, that for any invertible matrix $\boldsymbol{E}$, we have:

$$\frac{\partial E_{in}}{\partial E_{lm}^{-1}} = -E_{nm}E_{il} \tag{4.8.13}$$

Let $\boldsymbol{E}$ now be given by (4.2.14), so that:

$$\frac{\partial E_{mn}^{-1}}{\partial A_{ij}} = \sum_l \left(\delta_{jm} \cdot \left(\boldsymbol{COV}_{\vec{N}^{data}}^{-1}\right)_{il} A_{ln} + \delta_{jn} \cdot \left(\boldsymbol{COV}_{\vec{N}^{data}}^{-1}\right)_{ml} A_{li} \right) \tag{4.8.14}$$

Then we have

$$\frac{\partial E_{kl}}{\partial A_{ij}} = \sum_{mn} \frac{\partial E_{kl}}{\partial E_{mn}^{-1}} \cdot \frac{\partial E_{mn}^{-1}}{\partial A_{ij}} =$$

$$- E_{kj} \cdot \left(\boldsymbol{E} \cdot \boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right)_{li} - E_{jl} \cdot \left(\boldsymbol{E} \cdot \boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right)_{ki} \quad (4.8.15)$$

For $\vec{x}$, we obtain with (4.2.15):

$$\frac{\partial x_k}{\partial A_{ij}} = \sum \frac{\partial E_{kl}}{\partial A_{ij}} \cdot \left(\left[\boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \vec{N}^{data} + \tau^2 \cdot \boldsymbol{L}^2 \cdot \vec{x}^{Theory} \right] \right)_l$$

$$+ E_{kl} \cdot \frac{\partial}{\partial A_{ij}} \left(\left[\boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \vec{N}^{data} \right] \right)_l$$

And finally:

$$\frac{\partial x_k}{\partial A_{ij}} =$$

$$E_{kj} \cdot \left(\boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \vec{N}^{data} - \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \cdot \boldsymbol{A}\vec{x} \right)_i - \left(\boldsymbol{E} \cdot \boldsymbol{A}^T \cdot \boldsymbol{COV}_{\vec{N}^{data}}^{-1} \right)_{ki} \cdot \vec{x}_j \quad (4.8.16)$$

With (4.8.16) and (4.8.12), the uncorrelated systematic uncertainties $\Delta^{unc} x_i$ can be directly computed.

4.8.3   Treatment of Correlated Model Uncertainties

Correlated errors arise, whenever the entries N_{jk}^{MC} of the scaled unfolding matrix $\boldsymbol{N}^{MC}$ depend on some parameter η of the measurement, i.e. if we have a functional dependence for N_{jk}^{MC}:

$$\eta \rightarrow N_{jk}^{MC}(\eta) \quad (4.8.17)$$

For each such parameter η_λ we have uncertainties $\Delta\eta_{\lambda+}$ and $\Delta\eta_{\lambda-}$, which propagate to the entries of $\boldsymbol{N}^{MC}$. This way, we will obtain two new matrices $\boldsymbol{N}^{\lambda+}$ and $\boldsymbol{N}^{\lambda-}$ with shifted entries $N_{ij}^{\lambda+}$ and $N_{ij}^{\lambda-}$. The shifted unfolding matrices will alter the unfolding result establishing shifts $\Delta^{\lambda+} x_i$ and $\Delta^{\lambda-} x_i$ on every single data point x_i.

Typical parameters η on which the unfolding matrix depends are all sorts of calibration scales, since they often affect the whole kinematic range in a correlated manner. We will refer to the asymmetric uncertainties $\Delta\eta_{\lambda+}$ and $\Delta\eta_{\lambda-}$ as *input shifts*.

In the context of this analysis, we estimate the shifted matrix elements $N_{jk}^{\lambda\pm}$ using a simple Monte Carlo technique. The simulation is done independently for the three parameter values:

$$\eta_\lambda$$

$$\eta_{\lambda+} := \eta_\lambda + \Delta\eta_{\lambda+}$$

$$\eta_{\lambda-} := \eta_\lambda + \Delta\eta_{\lambda-}$$

They yield the matrices $\boldsymbol{N}^{MC}$, $\boldsymbol{N}^{\lambda+}$ and $\boldsymbol{N}^{\lambda-}$, respectively. For the systematic uncertainties $\Delta^{\lambda+} N_{jk}$ we then obtain simple estimators:

$$\Delta^{\lambda\pm} N_{jk} := N_{jk}^{\lambda\pm} - N_{jk}^{MC} \quad (4.8.18)$$

In order to further propagate the shifts $\Delta^{\lambda\pm}N_{jk}$ to the solution $\vec{x}$, we use multiple approaches. The package `TUnfoldSys` uses standard error propagation. From here on, this method is regarded to as *propagation method*:

$$\Delta^{\lambda\pm}x_i := \sum_{jk} \frac{\partial x_i}{\partial N_{jk}^{MC}} \cdot \Delta^{\lambda\pm}N_{jk}$$

$$\Delta^{\lambda\pm}x_i = \sum_{jklm} \frac{\partial x_i}{\partial A_{lm}} \cdot \frac{\partial A_{lm}}{\partial N_{jk}^{MC}} \cdot \Delta^{\lambda\pm}N_{jk} \tag{4.8.19}$$

Now, the correlated systematic uncertainties $\Delta^{\lambda\pm}x_i$ can directly computed with (4.8.16) and (4.8.10).

The *offset method* is even simpler, comparing the outcome of three independent unfolding processes with $\boldsymbol{N}^{MC}$, $\boldsymbol{N}^{\lambda+}$ and $\boldsymbol{N}^{\lambda-}$. The results $\vec{x}$, $\vec{x}^{\lambda+}$ and $\vec{x}^{\lambda-}$ are then compared, giving shifts $\Delta_{offset}^{\lambda\pm}x_i$ for each data point x_i:

$$\Delta_{offset}^{\lambda\pm}x_i := x_i - x_i^{\lambda\pm} \tag{4.8.20}$$

If the error propagation is reliable, offset and propagation method should approximately give the same result.

For comparison, a third method is studied, which is called the *bin-by-bin method* in this context. From independent Monte Carlo simulations for η_λ, $\eta_\lambda + \Delta\eta_{\lambda+}$ and $\eta_\lambda + \Delta\eta_{\lambda-}$ we compare the reconstructed distributions $\vec{N}^{rec}$, $\vec{N}^{\lambda+}$ and $\vec{N}^{\lambda-}$ and analyze the shifts on the data points:

$$\Delta_{bbb}^{\lambda\pm}x_i := \frac{\mathcal{L}_{data}}{\mathcal{L}_{MC}} \cdot \left(N_i^{rec} - N_i^{\lambda\pm} \right) \tag{4.8.21}$$

(Note the normalization condition in this definition.) This method basically neglects the presence of migration effects. Nevertheless, it is useful: By comparing its result to the other two methods, one can study the impact of migration effects on the systematic uncertainties.

For consistency reasons, the propagation method has been chosen to be quoted; the other two methods are merely used for comparison. Moreover, the systematic shifts will be symmetrized by the following *symmetrization prescription*:

$$\Delta_{sym}^{\lambda}x_i := \frac{1}{2}\left| \Delta^{\lambda+}x_i - \Delta^{\lambda-}x_i \right| \tag{4.8.22}$$

Note, that due to statistical fluctuation, $\Delta^{\lambda+}x_i$ and $\Delta^{\lambda-}x_i$ can sometimes have the same sign. This is not believed to be worrisome, since statistical fluctuations are accounted for by $\Delta^{stat}x_i$.

4.8.4 Regularization Uncertainty

By the nature of unfolding, the result $\vec{x}$ depends on the choice of the regularization parameter τ. Therefore, the error propagation from τ to $\vec{x}$ can be studied. We have for shifts $\Delta\tau_\pm$ (*propagation method*):

$$\Delta^{\tau\pm}x_i := \Delta\tau_\pm \cdot \frac{\partial x_i}{\partial\tau^2} \cdot 2\tau \tag{4.8.23}$$

Here, with (4.2.15) we have:

$$\frac{\partial x_i}{\partial \tau^2} = \left(\boldsymbol{E} \cdot \boldsymbol{L}^2 \left(\vec{x}^{Theory} - \vec{x} \right) \right)_i \tag{4.8.24}$$

We can also study the effect using the *offset method*, comparing the results $\vec{x}$, $\vec{x}^{\tau+}$ and $\vec{x}^{\tau-}$ for the different regularization parameters τ, $\tau + \Delta \tau_+$ and $\tau + \Delta \tau_-$. We define:

$$\Delta_{offset}^{\tau\pm} x_i := x_i - x_i^{\tau\pm} \tag{4.8.25}$$

If the error propagation is reliable, both results should give similar results.

While the propagation of the regularization parameter is well defined, it is much less clear how to estimate $\Delta \tau_\pm$. In fact, reasons can be found that justify to set $\Delta \tau_+ = \Delta \tau_- = 0$. This is, because a proper choice of τ should be conservative enough to not introduce any bias at all on $\vec{x}$. In this case, high values of $\frac{\partial}{\partial \tau^2} x_i$ rather reflect statistical fluctuations than a systematic uncertainty from regularization. Statistical uncertainties, however, are accounted for by $\Delta^{stat} x_i$.

Chapter 5

Cross Section Determination

The measurement of an inclusive cross section is at is core a *counting experiment.*
Given the data luminosity $\mathcal{L}_{data}$ and its measured event count N^{data}, the cross section
is in principle given via the simple formula $\sigma = N^{data}/\mathcal{L}_{data}$. In a real life scenario,
multiple intricacies have to be accounted for, such as the presence of background
sources, the correction of efficiency and migration effects and the treatment of photon
radiation. In the context of this analysis, all this is reflected by the five step procedure
pointed out in this subsection.

5.1 The Measurement Procedure

5.1.1 Binning

The first step is *binning.* For every event, the *inelasticity y* and the *squared momentum
transfer Q^2* are measured. The results are then binned in a two-dimensional bin grid
$\mathcal{G}_{det}$ with bins $\mathcal{B}^{det}_{i_1 i_2}$, where the two dimensions correspond to the variables y and Q^2.
Such, a *raw event count $N^{raw}_{i_1 i_2}$* is obtained for each bin $\mathcal{B}_{i_1 i_2}$. The details of the chosen
bin grid are discussed in chapter 10.

5.1.2 Background Subtraction

The second step is *background subtraction.* There are multiple background sources.
The most dominant source originates from pions in the SpaCal calorimeter, which
are misidentified as electrons. Within the context of this analysis, the background is
estimated by a Monte Carlo simulation using the package `PYTHIA`, see chapter 7. Let
$N^{bgr}_{i_1 i_2}$ be the event count of simulated background events in bin $\mathcal{B}_{i_1 i_2}$, then the *data
event count in Bin $\mathcal{B}_{i_1 i_2}$* is given by (4.6.1):

$$N^{data}_{i_1 i_2} := N^{raw}_{i_1 i_2} - \frac{\mathcal{L}_{data}}{\mathcal{L}_{bgr}} \cdot N^{bgr}_{i_1 i_2} \equiv N - f_{bgr} \cdot N^{bgr}_{i_1 i_2} \tag{5.1.1}$$

Here, $\mathcal{L}_{data}$ and $\mathcal{L}_{bgr}$ denote the luminosities of the data sample and the background
model and f_{bgr} is called the *background scale factor.*

5.1.3   Efficiency and Migration Correction

The third step is *unfolding.* At this step, the limited detector efficiency and migration effects are accounted for, giving the *unfolded event count $x_{i_1 i_2}$ for Bin $\mathcal{B}_{i_1 i_2}$:*

$$x_{i_1 i_2} := \left(\boldsymbol{A}^{-1} \left(\vec{N}^{data} \right) \right)_{i_1 i_2} \tag{5.1.2}$$

Here, $\boldsymbol{A}^{-1} \left(\vec{N}^{data} \right)$ denotes *symbolically* the result of the unfolding process. In fact, two different approaches have bin applied in this analysis, which are referred to as the *unfolding method* and the *bin-by-bin method*, see chapter 4. For the bin-by-bin method we simply obtain:

$$x_{i_1 i_2} := \frac{N_{i_1 i_2}^{gen}}{N_{i_1 i_2}^{rec}} \cdot N_{i_1 i_2}^{data} \tag{5.1.3}$$

Here, $N_{i_1 i_2}^{gen}$ and $N_{i_1 i_2}^{rec}$ denote the generated and reconstructed event count of the Monte Carlo sample, respectively. Note, that for the unfolding method, the bin grid $\mathcal{G}_{gen}$ for the unfolding result $\vec{x}$ typically differs from the bin grid $\mathcal{G}_{det}$ for the measured event counts $\vec{N}^{data}$, $\vec{N}^{raw}$ and $\vec{N}^{bgr}$.

5.1.4   Radiative Corrections

The fourth step is *radiative corrections.* As pointed out in section (2.5), the measured event count $\vec{N}^{data}$ does not only reflect the Born-level cross section, but rather contains higher order contributions from initial state radiation (ISR), final state radiation (FSR) and QED compton events (QEDC).

   The correction is given by a continuous function $\mathcal{R} \left(y, Q^2 \right)$ of y and Q^2, which relates the radiative and non-radiative cross section:

$$\frac{\partial^2}{\partial y \partial Q^2} \sigma^{NC,norad} \left(y, Q^2 \right) = \mathcal{R} \left(y, Q^2 \right) \cdot \frac{\partial^2}{\partial y \partial Q^2} \sigma^{NC,rad} \left(y, Q^2 \right) \tag{5.1.4}$$

In principle, this correction factor can be calculated directly vom QED, which has been done in [Spi93].

   Since this analysis is a counting experiment, we need bin wise correction factors $\mathcal{R}_{i_1 i_2}$ for every bin $\mathcal{B}_{i_1 i_2}$, which can be defined as follows:

$$\iint_{\mathcal{B}_{i_1 i_2}} \frac{\partial^2}{\partial y \partial Q^2} \sigma^{NC,norad} \left(y, Q^2 \right) dy dQ^2 = \mathcal{R}_{i_1 i_2} \cdot \iint_{\mathcal{B}_{i_1 i_2}} \frac{\partial^2}{\partial y \partial Q^2} \sigma^{NC,rad} \left(y, Q^2 \right) dy dQ^2 \tag{5.1.5}$$

We will call the factor $\mathcal{R}_{i_1 i_2}$ the *radiative correction factor* for Bin $\mathcal{B}_{i_1 i_2}$, and we have:

$$x_{i_1 i_2}^{Born} := \mathcal{R}_{i_1 i_2} \cdot x_{i_1 i_2} \tag{5.1.6}$$

   In order to quantify $\mathcal{R}_{i_1 i_2}$, we adopt a Monte Carlo approach according to [Spi92]. We compare two independent simulations with radiative effects beeing turned on and off, respectively. Then, for every bin $\mathcal{B}_{i_1 i_2}$, we obtain two event counts $N_{i_1 i_2}^{rad}$ and $N_{i_1 i_2}^{norad}$. Then, the correction factor $\mathcal{R}_{i_1 i_2}$ is simply given by:

$$\mathcal{R}_{i_1 i_2} = \frac{\mathcal{L}_{rad}}{\mathcal{L}_{norad}} \cdot \frac{N_{i_1 i_2}^{norad}}{N_{i_1 i_2}^{rad}} \tag{5.1.7}$$

Note, that often the radiative correction factor $\mathcal{R}_{i_1 i_2}$ is expressed as a relative difference $\delta_{i_1 i_2}$, where $\delta_{i_1 i_2}$ and $\mathcal{R}_{i_1 i_2}$ are related via:

$$\mathcal{R}_{i_1 i_2} = \frac{1}{1 + \delta_{i_1 i_2}} \tag{5.1.8}$$

In chapter (11) we will determine the factors $\mathcal{R}_{i_1 i_2}$ together with systematic uncertainties $\delta^{rad}\left(x_{i_1 i_2}^{Born}\right)$ on the corrected cross sections.

5.1.5   Cross Section Determination

The fifth step is the *cross section determination*. It accounts for the luminosity $\mathcal{L}_{data}$ of the studied data sample and the bin volumes $Vol_{i_1 i_2}$. For the *bin-averaged differential cross section in Bin $\mathcal{B}_{i_1 i_2}$* we obtain:

$$\left\langle \frac{\partial \sigma}{\partial y \partial Q^2} \right\rangle_{i_1 i_2} := \frac{x_{i_1 i_2}^{Born}}{Vol_{i_1 i_2} \cdot \mathcal{L}_{data}} \tag{5.1.9}$$

Here, the bin volumes are given by:

$$Vol_{i_1 i_2} = \left(\left(Q^2\right)_{up}^{i_1} - \left(Q^2\right)_{down}^{i_1} \right) \cdot \left(y_{up}^{i_2} - y_{down}^{i_2} \right) \tag{5.1.10}$$

In (5.1.10), the symbols $(Q^2)_{up}^{i_1}$, $(Q^2)_{down}^{i_1}$, $y_{up}^{i_2}$ and $y_{down}^{i_2}$ denote upper and lower bin boundaries of bin $\mathcal{B}_{i_1 i_2}$.

In section (6.1) the luminosity $\mathcal{L}_{data}$ is determined and a measure for its uncertainty $\delta\left(\mathcal{L}_{data}\right)$ is given.

The complete list of cross sections is given in appendix A.

5.2   Cross Section Uncertainties

In section 4.8 the uncertainty treatment of unfolding was discussed, both for statistical and for systematic error sources. We refer to it to the total unfolding error $\delta^{tot}(x_{ij}^{Born})$. The two error sources not accounted for by this procedure are the systematic uncertainty $\delta^{rad}(x_{i_1 i_2}^{Born})$ for the radiative corrections and the overall uncertainty σ^{lum} for the luminosity. These can be accounted for by quadratic summation. We define:

$$\delta^{tot+r}\left(x_{i_1 i_2}^{Born}\right) = \sqrt{\left(\delta^{tot}(x_{i_1 i_2}^{Born})\right)^2 + \left(\delta^{rad}(\mathcal{R}_{i_1 i_2})\right)^2} \tag{5.2.1}$$

$$\delta^{tot+r+l}\left(x_{i_1 i_2}^{Born}\right) = \sqrt{\left(\delta^{tot+r}(x_{i_1 i_2}^{Born})\right)^2 + \left(\delta\left(\mathcal{L}_{data}\right)\right)^2} \tag{5.2.2}$$

All these quantities are quoted in appendix A.

5.3   Bin center correction

Note, that formula (5.1.9) gives a cross section that is bin-averaged over bin $\mathcal{B}_{i_1 i_2}$. However, our final interest is rather directed to the functional value

$$\sigma_{i_1 i_2}^{diff} := \left. \frac{\partial^2 \sigma}{\partial y \partial Q^2}\left(Q^2, y\right) \right|_{\substack{Q^2 = Q_{i_1}^2 \\ y = y_{i_2}}} \tag{5.3.1}$$

of the measured cross section at some point $(Q^2_{i_1}, y_{i_2})$ in Bin $\mathcal{B}_{i_1 i_2}$. These *point-wise differential cross sections* can then directly be compared to any theory prediction $\left(\sigma^{Theory}\right)^{diff}_{i_1 i_2}$ at that point. Moreover, the triples $\left(Q^2_{i_1}, y_{i_2}, \sigma^{diff}_{i_1 i_2}\right)$ can be used for fitting the functional form of $(Q^2, y) \to \frac{\partial^2 \sigma}{\partial y \partial Q^2}(Q^2, y)$ in the context of a *PDF-fit*.

In this analysis point-wise differential cross sections (5.3.1) are obtained by exploiting the mean-value theorem. We start from a given theoretic prediction $(Q^2, y) \to \frac{\partial\left(\sigma^{Theory}\right)}{\partial y \partial Q^2}(Q^2, y)$ and find points (Q^2_i, y_j) such that:

$$Vol_{i_1 i_2} \cdot \frac{\partial^2\left(\sigma^{Theory}\right)}{\partial Q^2 \partial y}\left(Q^2_{i_1}, y_{i_2}\right) = \iint_{\mathcal{B}_{i_1 i_2}}\left(\frac{\partial^2\left(\sigma^{Theory}\right)}{\partial y \partial Q^2}\left(Q^2, y\right)\right) dQ^2 dy \qquad (5.3.2)$$

Then, we have

$$\left\langle \frac{\partial^2\left(\sigma^{Theory}\right)}{\partial y \partial Q^2}\right\rangle_{i_1 i_2} = \left.\frac{\partial^2\left(\sigma^{Theory}\right)}{\partial y \partial Q^2}\left(Q^2, y\right)\right|_{\substack{Q^2 = Q^2_{i_1} \\ y = y_{i_2}}} \qquad (5.3.3)$$

and for (5.3.1) we obtain the following relation:

$$\sigma^{diff}_{i_1 i_2} = \left\langle \frac{\partial \sigma}{\partial y \partial Q^2}\right\rangle_{i_1 i_2} = \frac{x^{Born}_{i_1 i_2}}{Vol_{i_1 i_2} \cdot \mathcal{L}_{data}} \qquad (5.3.4)$$

Of course, the position of the points $\left(Q^2_{i_1}, y_{i_2}\right)$ contains essential information about the measurement and has to be quoted together with the differential cross sections $\sigma^{diff}_{i_1 i_2}$ at that point.

While the mean value theorem ensures the existence of a point $\left(Q^2_{i_1}, y_{i_2}\right)$ fulfilling (5.3.2), the choice of $\left(Q^2_{i_1}, y_{i_2}\right)$ is far from unique. This is due to the two dimensions involved. Therefore, a choice has to be made. For y_{i_2} we choose the arithmetic bin center:

$$y_{i_2} := \frac{1}{2} \cdot \left(y^{i_2}_{up} + y^{i_2}_{down}\right) \qquad (5.3.5)$$

Then, $Q^2_{i_1}$ is defined via:

$$\left(\left(Q^2\right)^{i_1}_{up} - \left(Q^2\right)^{i_1}_{down}\right) \cdot \frac{\partial^2\left(\sigma^{Theory}\right)}{\partial Q^2 \partial y}\left(Q^2_{i_1}, y_{i_1}\right) = \int_{(Q^2)^{i_1}_{down}}^{(Q^2)^{i_1}_{up}}\left(\frac{\partial^2\left(\sigma^{Theory}\right)}{\partial y \partial Q^2}\left(Q^2, y_{i_2}\right)\right) dQ^2$$

$$(5.3.6)$$

Here, the monotony of the Q^2-dependance ensures the uniqueness of $Q^2_{i_1}$.

For the theoretical cross section, the functional form of `H1PDF2009` [Kre, A+09] has been chosen, which summarizes previous measurements from HERA I.

Chapter 6

Data Sample and Event Selection

In this section, the data sample of this analysis is defined. We discuss the run selection, the used subtrigger and the analysis level cuts.

6.1   Run Selection and Data Luminosity

In this analysis, the positron runs from 2006 and 2007 are studied, with the exception of the low energy runs at the end of the HERA run time in 2007. The run range of this period is 468531 to 500611. A *run selection* is applied, which excludes runs with low quality based on the following criteria.

Trigger Phase. For the trigger settings, different profiles exist, called trigger phases. For regular data taking, trigger phase 2 and 3 are used. The profiles 0 and 1 are used at the start of each HERA luminosity fill, while a beam steering procedure is executed. Runs from that time period cannot be trusted. Therefore, we require a minimal trigger phase of 2 for all runs in our sample.

Run Quality. Each run is flagged as either good (G), medium (M) or poor (P). The classification is done by means of the operational status of major subdetectors, such as the CJC, the LAr or the SpaCal. Moreover, short runs with luminosity $\mathcal{L}_{run} < 0.1$ nb^{-1} are classified as poor. Only runs classified as medium or good as used for this analysis.

Readout Status. Only those runs are selected, for which the event information from the relevant subdetectors has been read out. For this analysis, the relevant subdetectors are the SpaCal, the CJC1 and CJC2, the CIP, the LAr and the lumi system. Moreover, we require the trigger S3 to be up and running.

High Voltage Status. For safety reasons, the voltage of detector components can be lowered during a run. Therefore, the voltage is monitored and saved for each event. If the drop-off holds on for more than $\frac{1}{3}$ of the run, the full run is excluded. Otherwise, single events are excluded and the luminosity of the run is corrected accordingly.

The selection criteria described above are usually referred to as *Good Run Cut*s.

For this analysis, the luminosity is measured using QED Compton events, see section 3.2.4. Note, that one conventionally introduces a cut on the z Vertex position z_{Vtx} for the luminosity measurement:

$$-35 \text{ cm } < z_{Vtx} < 35 \text{ cm} \tag{6.1.1}$$

This is beneficial, since it keeps the satellite bunches from influencing the measured luminosity. Of course, the same cut has to bee applied on analysis level. For the run
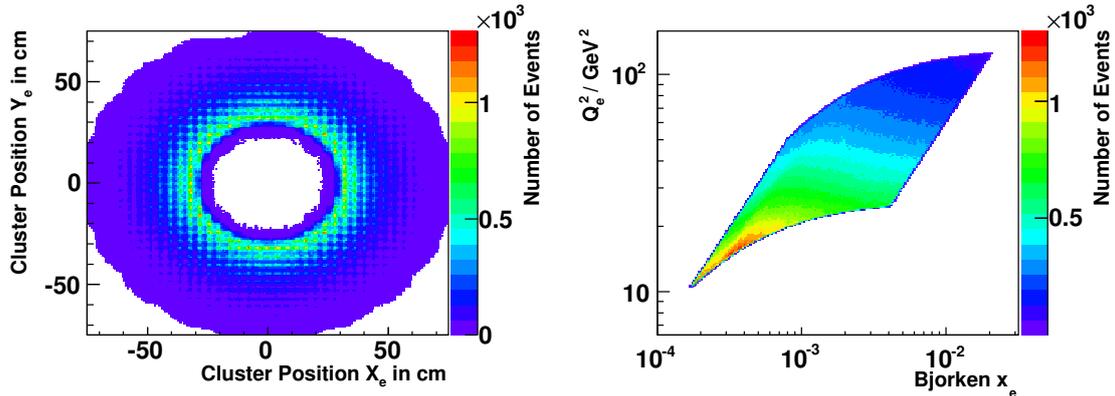
Figure 6.2.1: *Left: Data distribution as a function of the transverse SpaCal cluster position X_e and Y_e . The outer radial boundary of the SpaCal is clearly visible. Note, that the inner boundary reflects the radial trigger condition R > 30 cm. Right: Data distribution as a function of the Bjørken variable x_e and the squared momentum exchange Q_e^2. While the upper boundary corresponds to the outer geometric boundary of the SpaCal, the lower boundary reflects the radial trigger condition. The right boundary corresponds to the low inelasticity cut $y_e \geq 0.06$.*

selection discussed above, we obtain the following total integrated data luminosity:

$$\mathcal{L}_{data} \simeq 141.0 \text{ pb}^{-1} \tag{6.1.2}$$

Note, that the rather large systematic uncertainty of $\delta\left(\mathcal{L}_{data}\right) = 2\%$ is quoted for the luminosity measurement [SP].

6.2 Trigger Selection

For this analysis, the subtrigger S3 is used. It is sensitive to SpaCal clusters, which fulfill the following two *trigger conditions:*

$$E_e^{uncalib} > 10 \ GeV \tag{6.2.1}$$

$$R_e^{uncalib} > 30 \ cm \tag{6.2.2}$$

Here, $E_e^{uncalib}$ denotes the electromagnetic cluster energy and $R_e^{uncalib}$ the radial cluster position. Note, that both quantities are uncalibrated, since they are calculated at run time.

In figure 6.2.1, the data sample is plotted as a function of the transverse SpaCal cluster position X_e and Y_e. The outer radial boundary of the SpaCal is clearly visible. Note, that the inner boundary reflects the radial trigger condition $R_e^{uncalib} > 30$ cm.

Some of the subtriggers used in H1 use *prescales.* This is a prescription that aims at reducing the recording frequency of abundant event types. Thus, bandwidth is saved for rare event types which are typically of high scientific interest. Note however, that the subtrigger S3 is not prescaled.

The efficiency of the trigger conditions (6.2.1) and (6.2.2) will be discussed in chapter 8.

6.3 Analysis Level Event Selection

In this section the analysis level cuts are discussed.

6.3.1 Phase Space Cuts

We now formally define our phase space region of interest. Note that natural boundaries are given by the SpaCal geometry and the trigger conditions (6.2.1) and (6.2.2). This is reflected by introducing the following phase space cuts:

$$\textit{Energy Cut}: E_e \geq 11 \text{ GeV} \tag{6.3.1}$$

$$\textit{Radius Cut:} \ 30 \text{ cm} \leq R_e \leq 70 \text{ cm} \tag{6.3.2}$$

The final measurement will be done in the (Q_e^2, y_e)-plane. A rectangular area is defined by:

$$\textit{Inelasticity Cut:} \ 0.06 \leq y_e \leq 0.62 \tag{6.3.3}$$

$$\textit{Squared Momentum Exchange Cut}: 10 \text{ GeV}^2 \leq Q_e^2 \leq 133 \text{ GeV}^2 \tag{6.3.4}$$

In figure 6.2.1 the data sample is plotted. The left picture shows the distribution as a as a function of the SpaCal cluster position X_e and Y_e. The outer radial boundary of the SpaCal is clearly visible. The inner boundary reflects the radial trigger condition $R_e > 30$ cm. On the right side the data distribution is drawn as a function of the Bjørken variable x_e and the squared momentum exchange Q_e^2. While the upper boundary corresponds to the outer geometric boundary of the SpaCal, the lower boundary reflects the radial trigger condition. The right boundary corresponds to the low inelasticity cut $y_e \geq 0.06$.

6.3.2 z-Vertex Cut

For each event, we require the reconstruction of a primary vertex, which is needed to reconstruct a scattering angle θ_e for the electron. In order to suppress background from beam gas interactions and satellite bunches, we require the event vertices to lie close to the geometric center of the detector:

$$\textit{z-Vertex Cut:} \ -35 \text{ cm} \ < z_{Vtx} < 35 \text{ cm} \tag{6.3.5}$$

In figure 6.3.1 (left), the full z-Vertex distribution is shown. All events which fulfill (6.3.1), (6.3.2) and the trigger condition are binned. The complex longitudinal bunch structure consisting of the main bunch and the satellite bunches is clearly visible. The arrows indicate the boundaries of the z-Vertex cut.

Note, that we allow for two different vertex types to be reconstructed, which we refer to as *CJC vertex* and *CIP vertex*, see chapter 3. This is done to enhance the vertex reconstruction efficiency, as will be pointed out in chapter 8. In the case, that for an event both a CJC vertex and a CIP vertex is reconstructed, we prefer the CJC vertex.
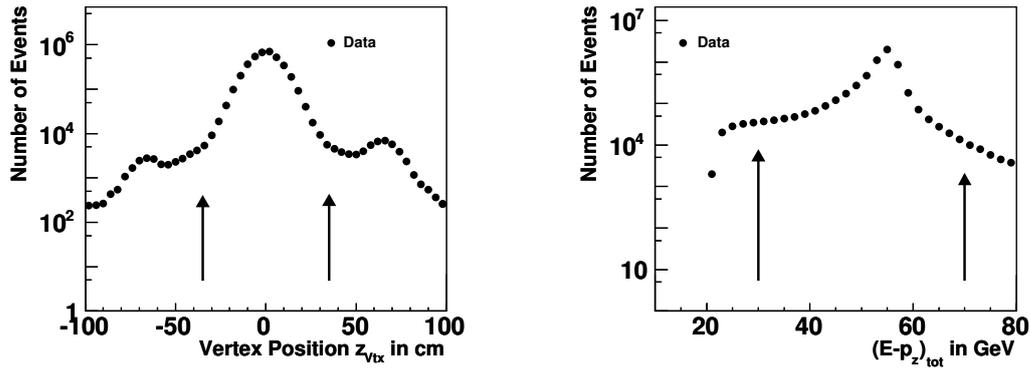
Figure 6.3.1: *Left: z-Vertex distribution with the main bunch and the two satellite bunches. Right: EMPZ-distribution with its peak around $2 \cdot E_{beam} \equiv 55.2$ GeV. For both plots, all cuts were released except the trigger condition and the cuts (6.3.1) and (6.3.2), which define the phase space under investigation. The arrows indicate the position of the cuts in each plot.*

6.3.3 The $E - p_z$ Cut

We also apply a cut on the following variable:

$$EMPZ \equiv (E - p_z)_{tot} = \sum_i (E - p_z)_i \tag{6.3.6}$$

The notation $EMPZ$ is used for brevity. Note, that the sum runs over all reconstructed particles. Therefore, the variable (6.3.6) describes the longitudinal energy momentum balance of the full final state. For non-radiative DIS events, we expect:

$$(E - p_z)_{tot} \simeq 2 \cdot E_{beam} \equiv 55.2 \text{ GeV} \tag{6.3.7}$$

By construction, the proton remant has $(E - p_z)_{remnant} = 0$, which makes the variable $EMPZ$ stable against particle losses into the forward direction.

However, $EMPZ$ is sensitive to radiative effects. ISR events have lower values for $EMPZ$ as the electron beam energy is effectively lowered. Even for FSR events, (6.3.7) is not necessarily fulfilled. To suppress radiative background, the following standard cut is used:

$$EMPZ \text{ Cut: } 30 \text{ GeV} \lesssim (E - p_z)_{tot} \lesssim 70 \text{ GeV} \tag{6.3.8}$$

Another type of background which is reduced by this cut is photoproduction ($Q^2 \simeq 0$). For this event type, typically no electron track is found, effectively reducing the reconstructed $EMPZ$-balance by its track momentum ($-p_{z,e} \simeq E_{beam} = 27.6$ GeV). Therefore, another cumulation in the $EMPZ$-distribution is expected around the kinematic peak ($E_e \sim 27.6$ GeV), which is removed by (6.3.8).

In figure 6.3.1 (right), the full $EMPZ$-distribution is shown for events that fulfill (6.3.1), (6.3.2) and the trigger condition. The peak around $2 \cdot E_{beam} \equiv 55.2$ GeV is clearly visible. The arrows indicate the boundaries of the $EMPZ$-cut.

6.3.4   Background Reduction

A potential source of background contamination are hadronic particles, which are wrongly reconstructed as electrons. This happens frequently for photoproduction events ($Q^2 \simeq 0$), for which the electron will be scattered into the beam pipe. The hadronic particles create sizable SpaCal clusters, which can then be mistaken as electrons. Note, that the hadronic background can originate from both charged and neutral tracks. Consider neutral pions for instance, to which the SpaCal is sensitive due to photo conversion ($\pi^0 \to \gamma\gamma$). In order to suppress this hadronic background, three more cuts are introduced.

Hadronic Energy Cut. Hadronic particles are expected to cause a significant energy deposit in the hadronic part of the SpaCal, which is typically larger than the one from electrons or photons. This can be used for further discrimination between DIS and photoproduction events. We require:

$$\textit{Hadronic Energy Cut: } E_{had} \leq 0.5 \text{ GeV} \tag{6.3.9}$$

In figure 6.3.3 (top right), distributions of the hadronic energy deposit E_{had} of electron clusters are shown for simulated DIS events (Djangoh) and simulated photoproduction events (Pythia). The arrow indicates the position of the cut.

ECRA Cut. Another variable, that exhibits discriminating power between data and background, is the cluster size $ECRA$, see chapter 5. Typically, hadronic particles leave a slightly larger cluster in the electromagnetic part of the SpaCal, suggesting a cut on this quantity. In figure 6.3.3 (bottom left), distributions of the electron cluster size $ECRA$ are shown for simulated DIS events (Djangoh) and simulated photoproduction events (Pythia).

A common value for a cut threshold is $ECRA_{tr} = 4$ cm. However, the distribution of the variable $ECRA$ is not very well described by the detector simulation, see figure 6.3.2 (top). Therefore, the inefficiency introduced by a cut on $ECRA$ is afflicted with a rather large systematic uncertainty. This problem can only be reduced by relaxing the cut. It follows, that the background reduction and the minimization of the systematic uncertainty are oppositional requirements for the $ECRA$ cut.

A solution to this problem is obtained by the observation, that the background contamination is limited to the region of small energy ($E_e \lesssim 16$ GeV) and small radius ($R_e \lesssim 55$ cm). Therefore, the $ECRA$ cut can be introduced as a function on E_e and R_e:

$$\textit{ECRA Cut: } ECRA \leq ECRA_{tr}\left(R_e, E_e\right) \tag{6.3.10}$$

Here, the *ECRA threshold function* $ECRA_{tr}\left(R_e, E_e\right)$ should vary within the interval $ECRA_{min} \lesssim ECRA_{tr}\left(R_e, E_e\right) \lesssim ECRA_{max}$ and must be constructed such, that the cut is relaxed for large radii and energies. The function is depicted in figure 6.3.4 (left). To formally define the function $ECRA_{tr}\left(R_e, E_e\right)$, we start with thresholds E_{min} and E_{max} for the energy and R_{min} and R_{max} for the radius. With the following helper functions

$$\xi_R\left(R_e, E_e\right) := ECRA_{min} + \frac{ECRA_{max} - ECRA_{min}}{R_{max} - R_{min}}\left(R_e - R_{min}\right) \tag{6.3.11}$$

$$\xi_E\left(R_e, E_e\right) := ECRA_{min} + \frac{ECRA_{max} - ECRA_{min}}{E_{max} - E_{min}}\left(E_e - E_{min}\right) \tag{6.3.12}$$
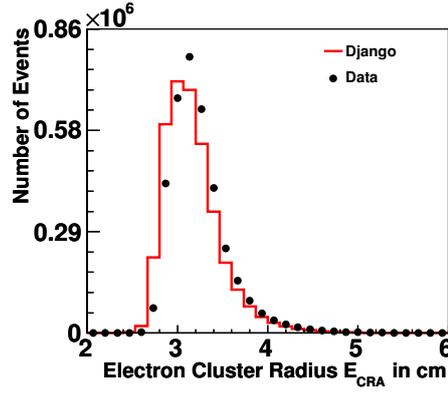
Figure 6.3.2: *ECRA distribution for data events and simulated events.*

$$\xi\left(R_e, E_e\right) := \max\left\{\xi_R\left(R_e, E_e\right), \xi_E\left(R_e, E_e\right)\right\} \tag{6.3.13}$$

we define:

$$ECRA_{tr}\left(R_e, E_e\right) = \begin{cases} ECRA_{max} & \text{if } E_e \geq E_{max}\text{or } R_e \geq R_{max} \\ ECRA_{min} & \text{if } E_e \leq E_{min}\text{and } R_e \leq R_{min} \\ \xi\left(R_e, E_e\right) & \text{otherwise} \end{cases} \tag{6.3.14}$$

For the ECRA interval we choose $ECRA_{min} = 3.8$ cm and $ECRA_{max} = 6$ cm. Note, that the variable ECRA never takes on values greater than 6 cm. Moreover, we choose $R_{min} = 50$ cm, $R_{max} = 55$ cm, $E_{min} = 14$ GeV and $E_{max} = 16$ GeV. The arrow in figure 6.3.3 (bottom left) indicates the position of the cut for small R_e and E_e. The ECRA cut introduces an inefficiency $\epsilon^{ECRA}\left(R_e, E_e\right)$, which will be discussed in chapter 8.

CIP Hit Cut. Another way to reduce background from neutral particles is provided by counting the CIP hits N_{hit}^{Cip} of the electron. This number is determined by searching for hits in a cone between the SpaCal and the primary vertex position. Since the electron is charged, it is expected to give multiple hits. Since the CIP detector has limited efficiency ϵ^{Cip}, we restrict the cut to the region with $E_e < 16$ GeV. This is the region of the phase space, which is susceptible to hadronic contamination. We require:

$$Cip\ Hit\ Cut: N_{hit}^{Cip} \geq N_{tr}^{Cip}\left(R_e, E_e\right) := \begin{cases} 2\forall E_e < 16 \text{ GeV} \\ 0\forall E_e > 16 \text{ GeV} \end{cases} \tag{6.3.15}$$

In figure 6.3.3 (bottom right), CIP hit counts are shown for simulated DIS events (Djangoh) and simulated photoproduction events (Pythia). The arrow indicates the position of the cut for small E_e. Effectively, the cut threshold $N_{tr}^{Cip}\left(R_e, E_e\right)$ is a function of the phase space variables, which is shown in 6.3.4 (right). The efficiency ϵ^{Cip} introduced by the Cip Hit Cut will be discussed in chapter 8.

The effect of the background reduction can be read off from figure 6.3.5, in which a part of the distribution of the electron energy E_e is shown for both data and simulated DIS events (Djangoh). The figures show the distributions before and after background reduction, respectively.
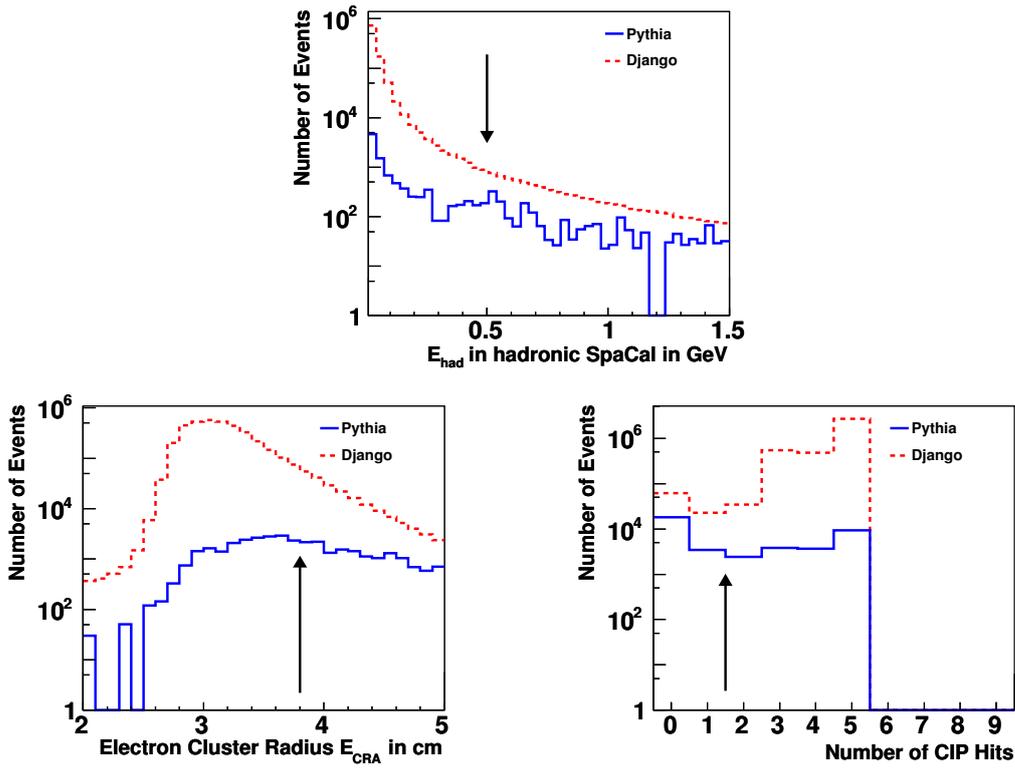
Figure 6.3.3: *Distributions of different background sensitive variables for simulated DIS events (Djangoh) and simulated photoproduction events (Pythia). Top Right: Hadronic energy deposit E_{had} of the reconstructed electron cluster. The arrow indicates the position of the applied cut. Bottom left: Electron cluster size ECRA. The arrow indicates position of the cut for small E_e and R_e. Bottom right: CIP hit count N_{hit}^{Cip}. The arrow indicates the position of the cut for small E_e.*
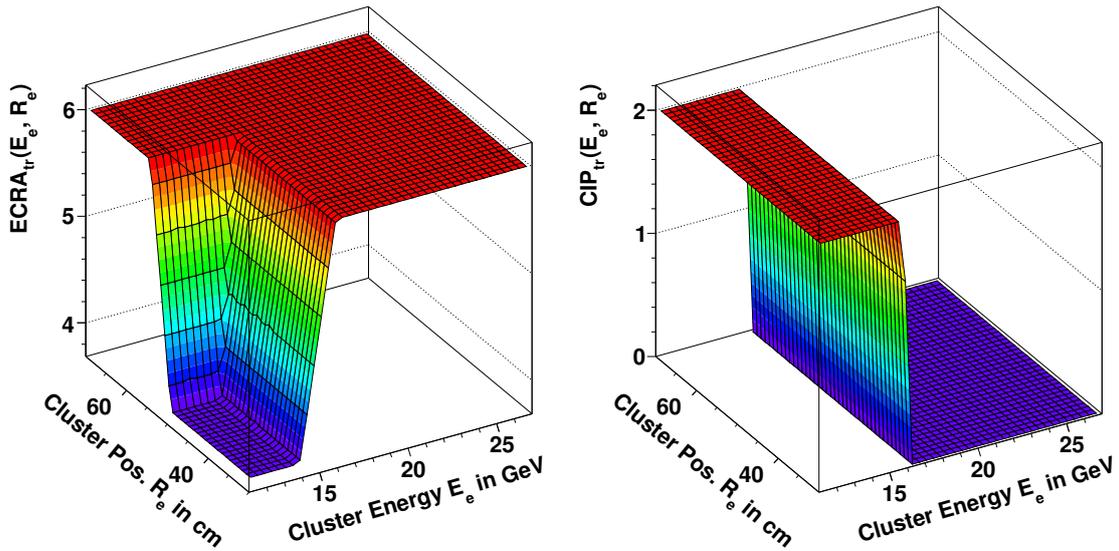
Figure 6.3.4: *Left: Threshold function $E_{tr}(R_e, E_e)$ for the ECRA cut according to (6.3.14). Right: Threshold function $N_{tr}^{Cip}(R_e, E_e)$ for the Cip Hit cut according to (6.3.15).*
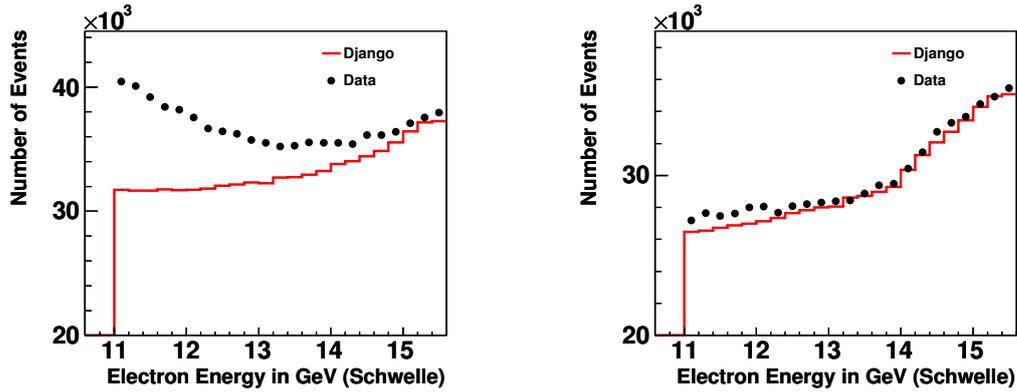
Figure 6.3.5: *Part of the distribution of the electron energy E_e for both data and simulated DIS events (Djangoh). Left: Before background reduction. Right: After background reduction according to (6.3.9), (6.3.10) and (6.3.15).*

6.3.5   The QEDC Cut

Another cut we apply to our sample is the exclusion of QED Compton events. This radiative event type is characterized by small momentum exchange between the lepton and the proton ($Q_{QED} \simeq 0$), see the discussion in chapter 2.

QED Compton events have a distinct signature, exhibiting two clusters which are back-to-back in azimuthal angle ϕ and fulfill $(E - p_z)_1 + (E - p_z)_2 = 2E_{beam}$. We therefore search for pairs of SpaCal clusters, which fulfill the following conditions:

$$-5° < \phi_1 - \phi_2 + 180° < 5° \tag{6.3.16}$$

$$-4 \text{ GeV} < (E - p_z)_1 + (E - p_z)_2 - 2 \cdot E_{beam} < 4 \text{ GeV} \tag{6.3.17}$$

$$7 \text{ GeV} < E_{1,2} < 22 \text{ GeV} \tag{6.3.18}$$

Events, which fulfill these conditions are likely to be QED Compton events and are excluded. In figure 6.3.6, a typical QED Compton event is shown with both clusters in the LAr calorimeter. In figure 6.3.7 the distribution of $(E - p_z)_1 + (E - p_z)_2$ and $\phi_1 - \phi_2$ are shown for the two reconstructed electron clusters with highest energy. The peak at $(E - p_z)_1 + (E - p_z)_2 \simeq 2 \cdot E_{beam} = 55.2$ GeV and $\phi_1 - \phi_2 \simeq \pi$ originating from QED Compton events is clearly visible. The arrows indicate the position of the applied cuts.
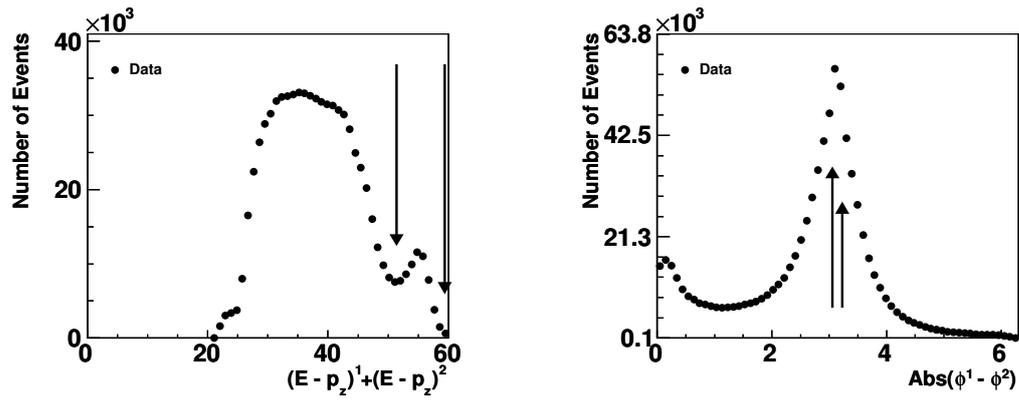
Figure 6.3.7: *Distribution of $(E - p_z)_1 + (E - p_z)_2$ (left) and $\phi_1 - \phi_2$ (right) for the two reconstructed SpaCal clusters with highest energy. For both plots, all cuts were released except the trigger condition and the cuts* (6.3.1) *and* (6.3.2). *The arrows indicate the range of the cuts.*
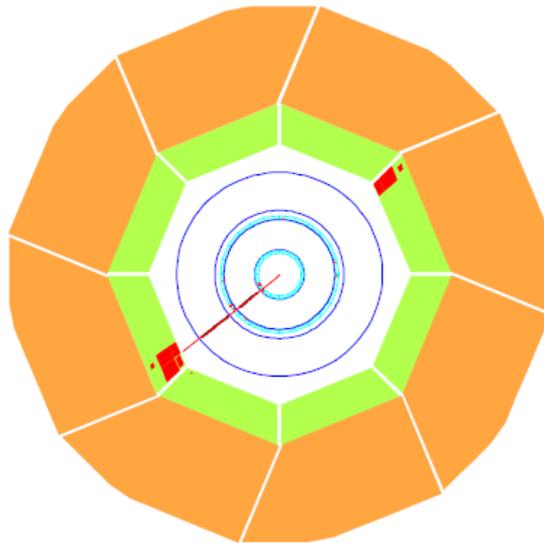
Figure 6.3.6: *Typical QED Compton event with two clusters in the LAr calorimeter. Both clusters are back-to-back in azimuth. The electron and the photon cluster can be distinguished by means of the reconstructed track.*

Chapter 7

The Monte Carlo Models

In the context of this analysis, two different Monte Carlo generators are used. The DIS events are modelled using the `DJANGOH` package while for the background simulation the `PYTHIA` package is used. The detector simulation and event reconstruction is done using the programs `H1SIM` and `H1REC`. In the first section of this chapter we introduce both generator models. In the second section we present control plots of different variables to study the consistency between the data and its Monte Carlo description.

7.1 The Djangoh Sample

The `DJANGOH` package [SS] generates DIS events including parton showers, hadronization and first order QED radiative corrections. It internally relies on `HERACLES` [KSM92] for the radiative corrections, `LEPTO` [IER97] for the hard scattering process, `ARIADNE` for parton showering and `JETSET` [S+01] for fragmentation and hadronization. `ARIADNE` is an implementation of the color dipole model [AGLP89, GP88] for parton cascades. `JETSET` relies on the Lund string model [AGIS83, AGS83]. For this work, the version 1.4 of `DJANGOH` has been used.

7.1.1 The Generator Sample

For this analysis 300 million events have been generated, accounting for radiative processes of ISR, FSR, QED Comption and and processes of Born level type. No generator weights have been applied to the events. The generation has been done using parton density functions from the set `CTEQ6L` [P+02, B+], which is accessible through `LHAPDF` [WBG05] (LHAPDF-Code: 10041). As kinematic cuts, $Q^2_{min} = 4$ GeV and $x_{min} = 10^{-5}$ have been used. Virtual corrections are applied to the QED vertex of the event. Integration over the phase space gives for the total generator cross section:

$$\sigma^{Djangoh}_{tot} = 347399 \text{ pb} \tag{7.1.1}$$

In the left column of figure 7.4.1 the generated distributions for the variables Q^2_e, y_e and x_e are drawn.

71

7.1.2   QCD Reweighting

The simulation should correspond to the H1 data rather than the to PDF set `CTEQ6L`. Therefore, a *QCD reweighting function* is applied to the Monte Carlo model, such that it effectively describes a sample according to the PDF set `H1PDF2009` [Kre, A$^+$09]. The weight function depends on the photon virtuality Q^2_{QCD} and the proton momentum fraction x_{QCD} carried by the struck parton. Both variables are available on generator level. The weight function is simply defined as the ratio of the double differential cross sections obtained from the two PDF sets:

$$w^{QCD}\left(x_{QCD}, Q^2_{QCD}\right) = \frac{\frac{\partial^2 \sigma^{H1PDF2009}}{\partial Q^2_{QCD} \partial x_{QCD}}\left(x_{QCD}, Q^2_{QCD}\right)}{\frac{\partial^2 \sigma^{CTEQ6L}}{\partial Q^2_{QCD} \partial x_{QCD}}\left(x_{QCD}, Q^2_{QCD}\right)} \tag{7.1.2}$$

Note, that this reweighting procedure needs to be applied both on detector level and on generator level. In figure 7.1.1 the weight function $w^{QCD}\left(x_{QCD}, Q^2_{QCD}\right)$ is depicted together with the structure functions $F_2^{H1PDF2009}(x, Q^2)$ and $F_2^{CTEQ6L}(x, Q^2)$ from which it is derived.

7.1.3   Djangoh Luminosity

In order to calculate the integrated luminosity of this Monte Carlo sample, we count all generated events passing the *Z Vertex Cut* (6.1.1) and the run selection from section 6.1 (*Good run cut*). Note, that for the luminosity determination the QCD weights w^{QCD} from (7.1.2) should not be applied to the events. However, the weights $w^{VtxRew}(z_{Vtx})$ for the Z vertex shape reweighting do have to be applied, see chapter 8. We end up with $N_{weighted} = 288.61$ million weighted events, or, equivalently:

$$\mathcal{L}_{model} = 830.8 \text{ pb}^{-1} \tag{7.1.3}$$

For further reference we introduce a model scale factor f_{model} for which we obtain with (6.1.2):

$$f_{model} = \frac{\mathcal{L}_{data}}{\mathcal{L}_{model}} = 0.1699 \tag{7.1.4}$$

From the data luminosity determination with QED Compton events the following systematic uncertainty is obtained [SP] :

$$\delta\left(f_{model}\right) = 2\% \tag{7.1.5}$$

We can recalculate the scale factor by comparing the sum of all weigthed reconstructed events for both the data and Monte Carlo distribution and obtain a slightly different result:

$$f_{model}^{shape} := \frac{\sum_i w_i^{data}}{\sum_i w_i^{model}} = 0.1648 \tag{7.1.6}$$

Here, each sum runs over all events selected on analysis level. The discrepancy between f_{model} and f_{model}^{shape} of the order of 3% is not compatible with the systematic uncertainty of the data luminosity. It rather reflects a discrepancy between the measurement of this analysis and `H1PDF2009`. We will come back to this issue in chapter 12.
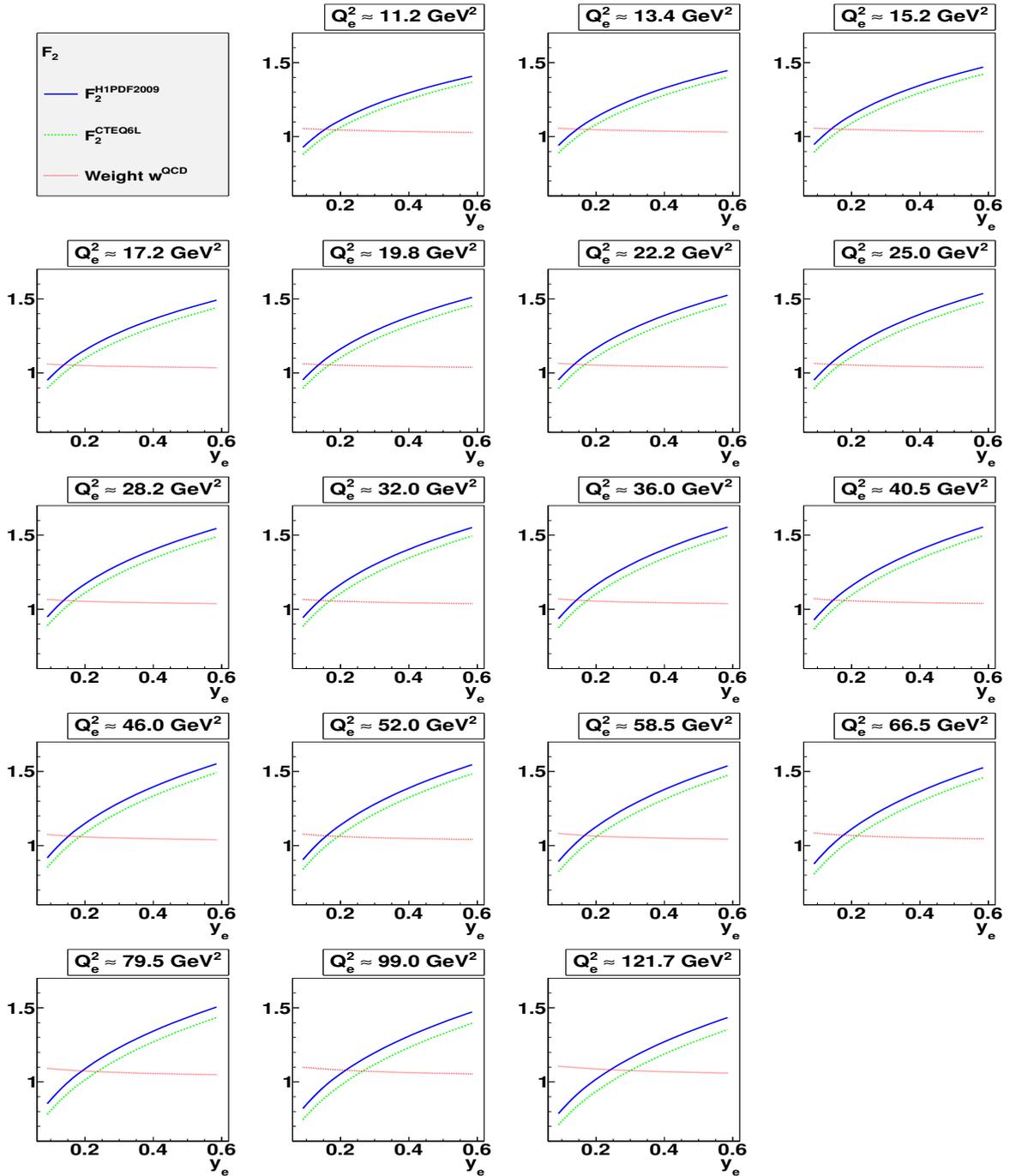
Figure 7.1.1: *Structure function $F_2(x, Q^2)$ for the pdf sets* H1PDF2009 [Kre, A$^+$09] *and* CTEQ6L [P$^+$02, B$^+$]. *Moreover, the QCD weight function $w^{QCD}\left(x_{QCD}, Q^2_{QCD}\right)$ is given.*

In figure 7.4.1 (left column) control plots of generated variables are given. Note the steep descent in the Q^2-distribution (top left) and the generator cut $Q^2 > 4$ GeV^2. Note the sharp peak in the y-distribution (center left) around $y \simeq 1$ which originates from ISR events, for which the center of mass energy is effectively lowered. Moreover, the distribution of the generated Bjorken variable x is shown (bottem left).

7.2 The Pythia Sample

The background is simulated using the `PYTHIA` [SMS08] package. A pure photoproduction sample is produced with $Q^2 \lesssim 4$ GeV. The generation has been done using parton density functions from the set `CTEQ5L` [L$^+$00] (`PDFLIB`-code: 1446 [PB93]). Integration over the phase space gives for the total generator cross section:

$$\sigma_{tot}^{Pythia} = 173176 \text{ pb} \tag{7.2.1}$$

In the right column of figure 7.4.1 the generated distributions for the variables Q_e^2, y_e and x_e are drawn.

We calculate the integrated luminosity of the Pythia sample in the same way we did for the Djangoh sample. Note, however, that the Pythia sample is afflicted with large generator weights w^{MC}. We end up with $N_{weighted} = 3.876$ million weighted events, or, equivalently:

$$\mathcal{L}_{bgr} = 223.85 \text{ pb}^{-1} \tag{7.2.2}$$

For further reference we introduce a background scale factor f_{bgr} for which we obtain with (6.1.2):

$$f_{model} = \frac{\mathcal{L}_{data}}{\mathcal{L}_{model}} = 0.63 \tag{7.2.3}$$

The determination of the background normalization suffers from large model uncertainty in this region of the phase space. However, the overall background contamination of the DIS sample is very low in the phase space of this analysis and, by the dedicated cuts described in chapter 6, is further reduced to values clearly below 1%. It follows that the estimation of the systematic unvertainty $\delta\left(\mathcal{L}_{bgr}\right)$ of the background luminosity is not very critical. Therefore, in accordance with [A$^+$03] the following rather conservative choice for the background scale uncertainty is made:

$$\delta\left(\mathcal{L}_{bgr}\right) = 0.5 \tag{7.2.4}$$

In figure 7.4.1 (right column) control plots of generated variables are given. The generated Q^2-variables are very low ($Q^2 \lesssim 2$ GeV^2), as photoproduction events are produced (top right). Moreover, the events exhibit high elasticity, see the y_e-distribution (center right). Moreover, the distribution of the generated Bjorken variable x is shown (bottom left).

7.3 Simulation and Reconstruction

The detector simulation is done with the package `H1SIM` which is based on `GEANT` [BBM$^+$]. It simulates the passing of the particles through the detector material including multiple scattering. Moreover, a calorimetric shower simulation is available.

The reconstruction is done in the same way as for genuine data events.

7.4   Control Plots of Reconstructed Variables

In figures 7.4.2 and 7.4.3, control plots of different variables are given to study the consistency between the data and its Monte Carlo description. The Djangoh sample is shape normalized according to (7.1.6).

In 7.4.2 (top left) the Q_e^2-dependence with its typical falling behaviour is shown. The lower limit is an artefact of the radius cut $R_e > 30$ cm. The y_e-distribution is shown in 7.4.2 (top right), while the logarithmic binning masks the descending behaviour which is typical for this variable. Note the kink in the top right corner, which originates from the ECRA Cut and the EMPZ Cut which are especially restrictive at small electron energies $E_e \lesssim 16$ GeV.

In 7.4.2 (center left), the Bjorken variable x_e is shown. One can easily understand the shape of the distribution by projection of figure 6.2.1 (right) on the horizontal axis. 7.4.2 (center right) shows the z-Vertex distribution exhibiting the Gaussian shape of the main bunch. Note, that events featuring either a CIP vertex, a CJC vertex or both are binned.

Figure 7.4.2 (bottom left) shows the well described polar angle distribution of the electron (θ_e). It shows the expected rise towards small scattering angles ($\theta_e \to \pi$). Note in figure 7.4.2 (bottom right) the S-shaped ascent of the energy distribution between 14 GeV and 16 GeV. It is an artifact of the ECRA cut and the EMPZ cut, which is are especially restrictive at small electron energies $E_e \lesssim 16$ GeV.

The plot for the variable $EMPZ \equiv (E - p_z)_{tot}$ is drawn in figure 7.4.3 (top left). A slight shift between the distributions can be observed. Recall, that a cut is applied on this variable, see (6.3.8). However, since the cut affects only the tails of the distribution, no efficiency discrepancy is expected to emerge from this cut.

The SpaCal cluster size $ECRA$ exhibits a severe shape discrepancy between data and Monte Carlo, see figure 7.4.3 (top right). Recall, that a cut is applied on this variable, see (6.3.10), which will affect the reconstruction efficiency ϵ^{Ecra}. In particular, due to the shape discrepancies in the $ECRA$-distribution, the reconstruction efficiencies for data ϵ_{data}^{Ecra} and model ϵ_{model}^{Ecra} will differ. Therefore, both efficiencies have to be determined carefully and an appropriate correction needs to be applied to the Monte Carlo simulation. This is done in chapter 8.

In figure 7.4.3 (bottom left) the number of vertex fitted tracks N_{tr} is plotted. Note, that the vertex fitting is done in reference to the CJC vertex. Events which do not have a vertex fitted track (and hence no CJC vertex) can still enter the event selection by featuring a CIP vertex. In figure 7.4.3 (bottom right) the distribution of the number of CIP hits N_{hits}^{Cip} is shown.

Control plots of the `PYTHIA` sample are shown in figure 7.4.4. The control plots are normalized to the data luminosity according to (7.2.3). Note the low event frequency for the background events. They predominantly occur at high y_e and small Q_e^2 values, which corresponds to low energy events ($E_e \lesssim 16$ GeV) at small scattering angles ($R_e \lesssim 50$ cm).
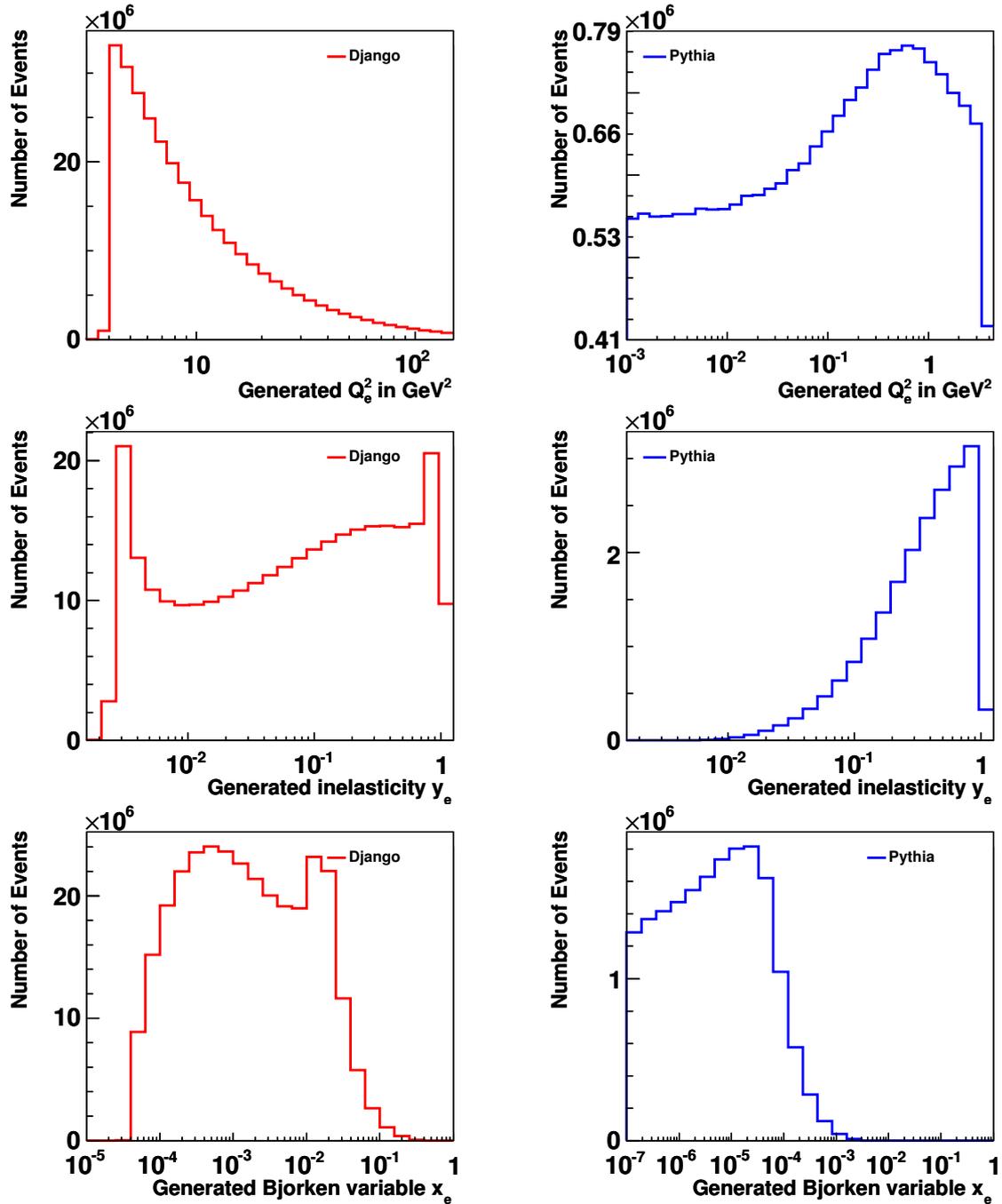
Figure 7.4.1: *Generator level distributions for the* DJANGOH *and the* PYTHIA *sample. Top: Squared momentum exchange* Q_e^2. *Center: Inelasticity* y_e. *Bottom: Bjorken variable* x_e.
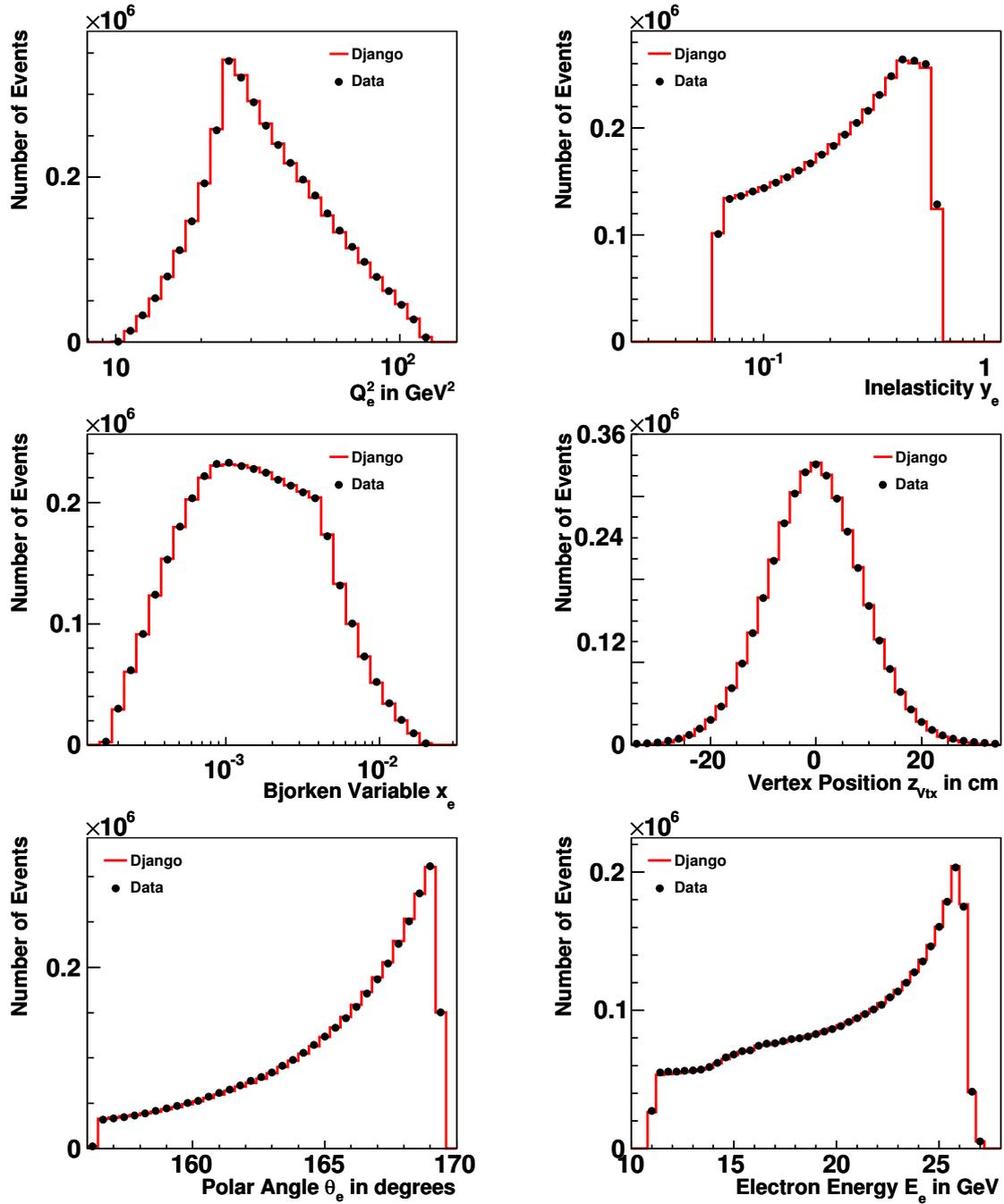
Figure 7.4.2: *Control plots of reconstructed variables for data and model (*DJANGOH*). Top: Squared momentum exchange Q_e^2 and Inelasticity y_e. Center: Bjorken variable x_e and z vertex position z_{Vtx}. Bottom: Electron polar angle θ_e and electron energy E_e.*
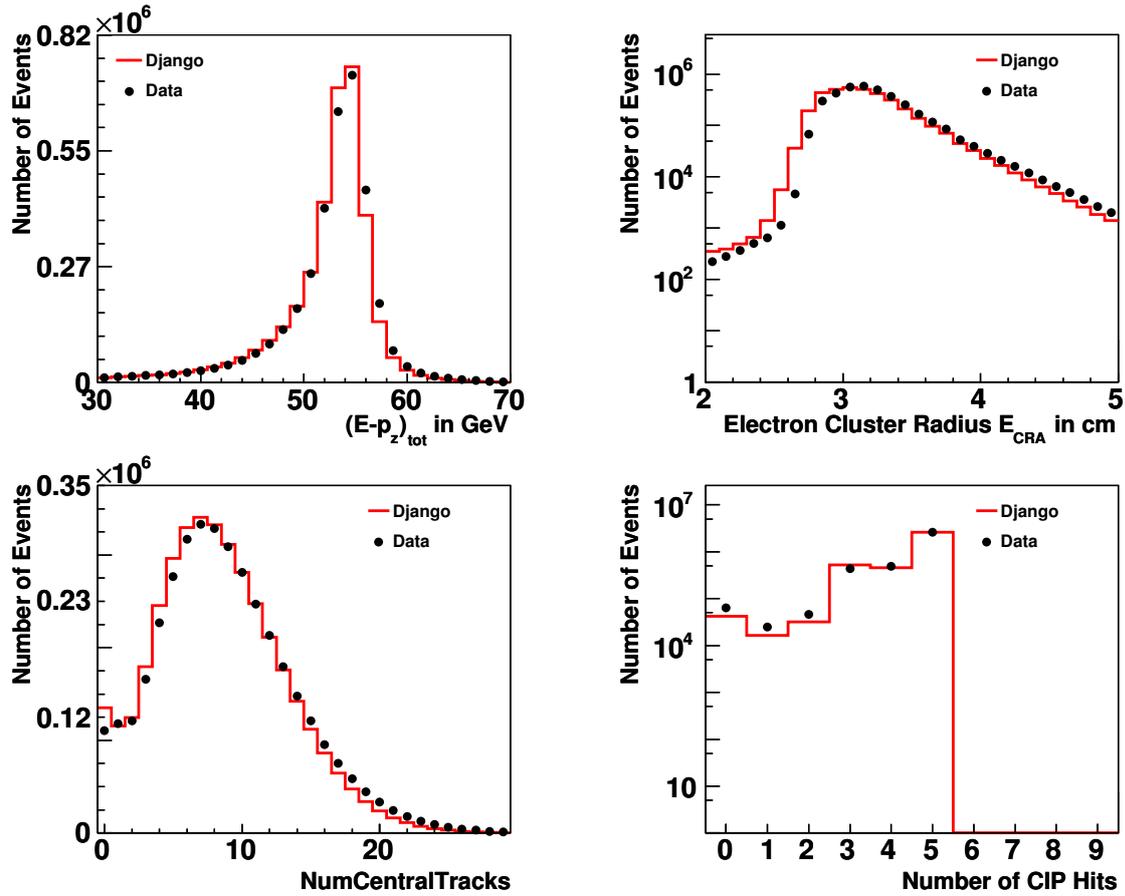
Figure 7.4.3: *Control plots of reconstructed variables for data and model (*DJANGOH*). Top: The variable $EMPZ \equiv (E - p_z)_{tot}$ and the SpaCal cluster size ECRA. Bottom: Number N_{tr} of reconstructed vertex fitted tracks and number N_{hit}^{Cip} of CIP hits along the electron track.*
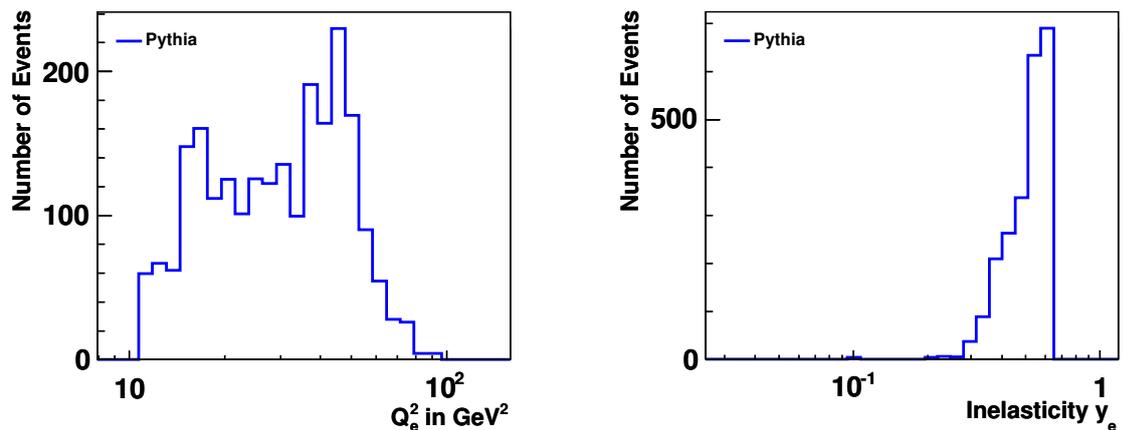
Figure 7.4.4: *Control plots of the reconstructed variables Q_e^2 and y_e for the background simulation (*PYTHIA*).*

Chapter 8

Efficiency and Z Vertex Correction

As pointed out in chapter 4, the success of the unfolding strategy critically depends on a reliable modelling of the detector effects. In this chapter, efficiency studies are performed for the data sample and for the Monte Carlo. Discrepancies between data and model efficiency are then remedied by the introduction of appropiate weight functions.

Another correction that needs to be applied to the Monte Carlo simulation is the z-Vertex reweighting. It will be discussed in section 8.3.

8.1   General Approach to Efficiency Correction

The reason for a potential loss of reconstruction efficiency is typically some sort of cut on the event sample for background reduction. For each such cut λ, we determine both a data efficiency $\epsilon^{\lambda,data}(\Xi_\lambda)$ and a model efficiency $\epsilon^{\lambda,model}(\Xi_\lambda)$. Here, both efficiencies are functions on some variable Ξ_λ, which is specific to the cut λ under investigation. For instance, the CIP efficiency $\epsilon^{Cip}(Z_{CIP})$ depends on the longitudinal coordinate Z_{CIP} of the electron's passing through the CIP layers. Note, that in some cases, the efficiency even depends on more than one variable. For such cases, the symbol Ξ_λ shall denote an n-tupel of variables, i. e. $\Xi_\lambda = (\Xi_\lambda^1, \Xi_\lambda^2, \ldots)$. For instance, as we will see in section 8.2, the efficiency of the ECRA cut is a function of the cluster position R_e and the cluster energy E_e.

In order to obtain the efficiencies $\epsilon^\lambda(\Xi_\lambda)$, background free reference samples need to be defined. Then, two independent analyses of the reference sample are performed, once with and once without the cut λ that introduces the inefficiency. We obtain distributions $N_{cut}^\lambda(\Xi_\lambda)$ and $N_{all}^\lambda(\Xi_\lambda)$, respectively. Note, that $N_{cut}^\lambda(\Xi_\lambda)$ and $N_{all}^\lambda(\Xi_\lambda)$ are not smooth functions, since they originate from the discrete data points of a counting experiment. Rather, $N_{cut}^\lambda(\Xi_\lambda)$ and $N_{all}^\lambda(\Xi_\lambda)$ can be viewed as step functions over an appropriate binning $\mathcal{G}_\lambda$ of Ξ_λ. The choice of the bin grid $\mathcal{G}_\lambda$ should reflect the available event statistics as well as the shapes of the efficiencies to be compared.

The efficiency is now simply the ratio between the two event distributions:

$$\epsilon^\lambda(\Xi_\lambda) = \frac{N_{cut}^\lambda(\Xi_\lambda)}{N_{all}^\lambda(\Xi_\lambda)} \tag{8.1.1}$$

while its statistical uncertainty is given by:

$$\Delta\epsilon^\lambda(\Xi_\lambda) = \sqrt{\frac{\epsilon^\lambda(\Xi_\lambda) \cdot (1 - \epsilon^\lambda(\Xi_\lambda))}{N_{all}^\lambda(\Xi_\lambda)}} \qquad (8.1.2)$$

This procedure is done for both the model and data efficiency.

8.1.1   Weight Functions

With $\epsilon^{\lambda,data}(\Xi_\lambda)$ and $\epsilon^{\lambda,model}(\Xi_\lambda)$, a *weight function* $w^\lambda(\Xi_\lambda)$ can be defined for each efficiency effect λ:

$$w^\lambda(\Xi_\lambda) = \frac{\epsilon^{\lambda,data}(\Xi_\lambda)}{\epsilon^{\lambda,model}(\Xi_\lambda)} \qquad (8.1.3)$$

Note, that unlike the efficiencies $\epsilon^{\lambda,data}(\Xi_\lambda)$ and $\epsilon^{\lambda,model}(\Xi_\lambda)$, the weight function $w^\lambda(\Xi_\lambda)$ can take on values both greater and less than 1. For the cases studied in section 8.2, all weights are indeed close to 1, see figures 8.2.2, 8.2.4 (left), 8.2.7, 8.2.9 and the discussion there.

With the weight functions $w^\lambda(\Xi_\lambda)$ at hand, the Monte Carlo simulation can be redone, such that all efficiencies are modelled correctly. To achieve this, every single Monte Carlo event is applied the following weight on detector level:

$$W = \prod_\lambda w^\lambda(\Xi_\lambda) \qquad (8.1.4)$$

Here, the product runs over all sources λ of an inefficiency. This way, an efficiency corrected detector matrix $\tilde{\boldsymbol{N}}^{MC}$ and an efficiency corrected reconstructed Monte Carlo distribution $\tilde{\vec{N}}^{rec}$ are obtained:

$$\boldsymbol{N}^{MC} \rightarrow \tilde{\boldsymbol{N}}^{MC}$$

$$\vec{N}^{rec} \rightarrow \tilde{\vec{N}}^{rec}$$

Of course, no weights are applied for the generator level distribution $\vec{N}^{gen}$, which means we have $\tilde{\vec{N}}^{gen} \equiv \vec{N}^{gen}$. Then, by construction, we have $\tilde{\epsilon}^{\lambda,model} \simeq \epsilon^{\lambda,data}$ for the re-evaluated model efficiency.

Evidently, the determination of the weights $w^\lambda(\Xi_\lambda)$ depends on the reliability of the efficiency determination and the available event statistics. This is, why for some regions in Ξ_λ weights cannot be safely determined. For instance, the CIP efficiency reweighting $w^{Cip}(Z_{CIP})$ is only possible within regions with efficiency significantly greater than 0. Outside these region, the weights will therefore be fixed to the boundary values.

Formally, we refine the definition of the weight functions by introducing a *region of trust* $\Omega_\lambda \equiv [\Xi_\lambda^{min}, \Xi_\lambda^{max}]$ for each $w^\lambda(\Xi_\lambda)$. For $\Xi_\lambda \notin \Omega_\lambda$ we regard $w^\lambda(\Xi_\lambda)$ as not reliable and use $w^\lambda(\Xi_\lambda^{min})$ or $w^\lambda(\Xi_\lambda^{max})$ instead. We define:

$$\hat{w}^\lambda(\Xi_\lambda) = \begin{cases} w^\lambda(\Xi_\lambda^{min}) & \forall \Xi_\lambda < \Xi_\lambda^{min} \\ w^\lambda(\Xi_\lambda) & \forall \Xi_\lambda \in \Omega_\lambda \\ w^\lambda(\Xi_\lambda^{max}) & \forall \Xi_\lambda > \Xi_\lambda^{max} \end{cases} \qquad (8.1.5)$$

This definition is particularly useful to exclude regions of low efficiency $\epsilon^\lambda \ll 1$ from the region of trust. In those areas large systematic uncertainties on the weight functions would emerge. From now on, with the symbol $w^\lambda(\Xi_\lambda)$, we *implicitly* understand a weight function according to (8.1.5).

Note, that the weight function $w^\lambda(\Xi_\lambda)$ for some cut λ is not given in a functional form $\Xi_\lambda \to w^\lambda(\Xi_\lambda)$, but rather as a set of bin wise data points $w_1^\lambda, w_2^\lambda, w_3^\lambda$ ... Therefore, for each efficiency study λ an appropriate binning $\mathcal{G}_\lambda$ of the variable Ξ_λ must be chosen. This binning $\mathcal{G}_\lambda$ should reflect the shape of the observed weight function $w^\lambda(\Xi_\lambda)$ and the available statistics of the reference sample.

Unless specified otherwise, from here on the symbol $\boldsymbol{N}^{MC}$ shall implicitly denote the *reweighted* response matrix.

8.1.2 Estimation of Systematic Uncertainties

The idea to account for the systematic uncertainty of the efficiency correction is the introduction of *shifted weight functions* $w^{\lambda+}(\Xi_\lambda)$ and $w^{\lambda-}(\Xi_\lambda)$ for each efficiency effect λ. For instance, we shift the CIP efficiency weight function $w^{Cip}(Z_{CIP})$ up and down, giving $w^{Cip+}(Z_{CIP})$ and $w^{Cip-}(Z_{CIP})$. Then, the Monte Carlo simulation is repeated for both of these shifted weight functions, giving two shifted response matrices $\boldsymbol{N}^{Cip+}$ and $\boldsymbol{N}^{Cip-}$. These matrices can then be used for the error propagation according to section 4.8.

More formally, we produce two independent Monte Carlo simulations for each λ, while each event is applied the following weight on detector level:

$$W^{\lambda\pm} = \prod_\eta w^\eta(\Xi_\eta) \times \frac{w^{\lambda\pm}(\Xi_\lambda)}{w^\lambda(\Xi_\lambda)} \tag{8.1.6}$$

This means effectively, that for the efficiency effect λ under investigation, only the weight function $w^\lambda(\Xi_\lambda)$ is shifted ($w^\lambda(\Xi_\lambda) \to w^{\lambda\pm}(\Xi_\lambda)$), while for each $\eta \neq \lambda$ no weight shifting is performed. With this procedure, two shifted response matrices $\boldsymbol{N}^{\lambda\pm}$ and two shifted detector level distributions $\vec{N}^{rec,\pm\Delta_\lambda}$ are obtained for each λ:

$$\boldsymbol{N}^{MC} \to \boldsymbol{N}^{\lambda\pm}$$

$$\vec{N}^{rec} \to \vec{N}^{rec,\lambda\pm}$$

The shifts $\vec{N}^{rec,\lambda\pm} - \vec{N}^{rec}$ can then be used as estimates of the systematic uncertainties on $\vec{N}^{rec}$ originating from the efficiency effect λ. With the shifted matrices $\boldsymbol{N}^{\lambda\pm}$ at hand, the error propagation of these uncertainties to the unfolding result $\vec{x}$ can be performed, using the techniques described in section 4.8.

The remaining task is the definition of the shifted weight functions $w^{\lambda\pm}(\Xi_\lambda)$. This is done by shifting the model efficiency $\epsilon^{\lambda,model}(\Xi_\lambda)$ up and down, giving $\epsilon_+^{\lambda,model}(\Xi_\lambda)$ and $\epsilon_-^{\lambda,model}(\Xi_\lambda)$, which propagates to the weight functions $w^{\lambda+}(\Xi_\lambda)$ and $w^{\lambda-}(\Xi_\lambda)$ via (8.1.3). For most efficiency effects λ, we will take the difference between the data and the model efficiency as a measure for the systematic uncertainty of the model efficiency, i.e. $\Delta^{sys}\left(\epsilon^\lambda(\Xi_\lambda)\right) = \Delta\epsilon^{\lambda,data}(\Xi_\lambda) - \Delta\epsilon^{\lambda,model}(\Xi_\lambda)$. Therefore, for the shifted model efficiencies $\epsilon_\pm^{\lambda,model}(\Xi_\lambda)$, the following choice is motivated:

$$\epsilon_\pm^{\lambda,model}(\Xi_\lambda) = \max\left\{\epsilon^{\lambda,model}(\Xi_\lambda) \pm \frac{1}{2} \cdot \left(\epsilon^{\lambda,data}(\Xi_\lambda) - \epsilon^{\lambda,model}(\Xi_\lambda)\right), 1\right\}$$

Here, the maximum prescription ensures, that the shifted model efficiencies never take on values greater than 1.

More general, we define the shifted model efficiencies $\epsilon_{\pm}^{\lambda,model}(\Xi_\lambda)$ in the following way:

$$\epsilon_{\pm}^{\lambda,model}(\Xi_\lambda) := \max\left\{\epsilon^{\lambda,model}(\Xi_\lambda) \pm f^\lambda \cdot \left(\epsilon^{\lambda,data}(\Xi_\lambda) - \epsilon^{\lambda,model}(\Xi_\lambda)\right) \pm p^\lambda, 1\right\} \quad (8.1.7)$$

Note, that both $\epsilon_{+}^{\lambda,model}(\Xi_\lambda)$ and $\epsilon_{-}^{\lambda,model}(\Xi_\lambda)$ can take on values greater or less than $\epsilon^{\lambda,model}(\Xi_\lambda)$, but never greater than 1. Here, f^λ (*shift factor*) and p^λ (*shift pedestal*) are parameters, that can be used to adjust the level of shifting. As a standard value, we use $f^\lambda = 0.5$ and $p^\lambda = 0$, which corresponds to the choice motivated above.

We now define the shifted weight functions in analogy to (8.1.3):

$$w^{\lambda\pm}(\Xi_\lambda) := \frac{\epsilon^{\lambda,data}(\Xi_\lambda)}{\epsilon^{\lambda\pm}(\Xi_\lambda)} \quad (8.1.8)$$

Again, we refine this notion by limiting this weighting prescription to the region of trust $\Omega_\lambda \equiv [\Xi_\lambda^{min}, \Xi_\lambda^{max}]$ and define:

$$\hat{w}^{\lambda\pm}(\Xi_\lambda) = \begin{cases} w^{\lambda\pm}(\Xi_\lambda^{min}) & \forall \Xi_\lambda < \Xi_\lambda^{min} \\ w^{\lambda\pm}(\Xi_\lambda) & \forall \Xi_\lambda \in \Omega_\lambda \\ w^{\lambda\pm}(\Xi_\lambda^{max}) & \forall \Xi_\lambda > \Xi_\lambda^{max} \end{cases} \quad (8.1.9)$$

From now on, with the symbol $w^{\lambda\pm}(\Xi_\lambda)$, we *implicitly* understand a weight function according to (8.1.9).

8.2 Determination of Efficiencies

In this subsection, we discuss the S3 trigger efficiency, the CIP efficiency, the Zvtx reconstruction efficiency and the ECRA cut efficiency.
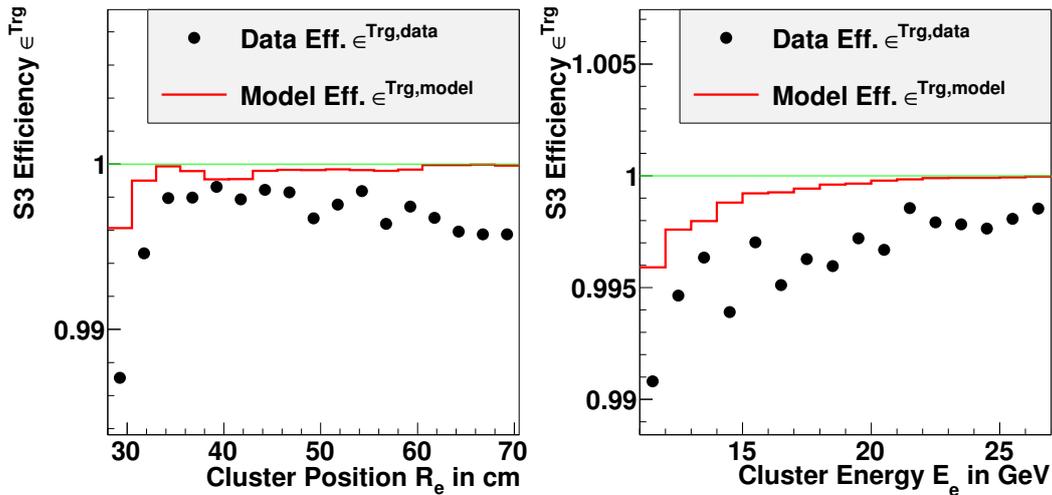
Figure 8.2.1: *S3 trigger efficiencyϵ^{Trg}. Left: As a function of the radial position R_e of the SpaCal electron. Right: As a function of the electromagnetic energy E_e of the SpaCal electron.*
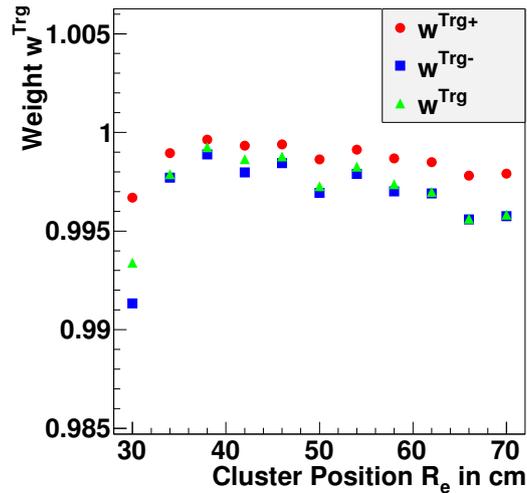
Figure 8.2.2: *S3 trigger efficiency weight functions w^{Trg}, w^{Trg+} and w^{Trg-} as a function of the radial position R_e of the SpaCal electron.*

8.2.1   S3 Trigger efficiency

In order to determine the S3 trigger efficiency, an independent reference sample has been defined using the S74 trigger. While the S3 trigger is sensitive to the SpaCal electron clusters, S74 triggers dijet events in the LAr calorimeter. Therefore, no correlation between the two triggers is expected.

In figure 8.2.1 the trigger efficiency is plotted, once as function of the SpaCal position R_e of the cluster and once as a function of its energy E_e. A dropping of the data efficiency towards the radius threshold $R_{tr} \simeq 30$ cm and towards the energy threshold at $E_{tr} \simeq 10$ GeV is visible. A slight difference between $\epsilon^{Trg,data}$ and $\epsilon^{Trg,model}$ is visible, that varies with R_e. Its order of magnitude is about 0.5 %.

The reweighting $w^{Trg}(R_e)$ is applied as a function of R_e, since the efficiency difference varies prodominantly with R_e. For the reweighting procedure, a rather course bin grid $\mathcal{G}_{Trg}$ of 11 bins between 28 and 72 cm has been chosen. For the systematic uncertainties, we set $f^{Trg} = 0.5$ and $p^{Trg} = 0$. The result is shown in figure 8.2.2, where w^{Trg}, w^{Trg+} and $w^{Trg-}$are depicted for the bins of $\mathcal{G}_{Trg}$.
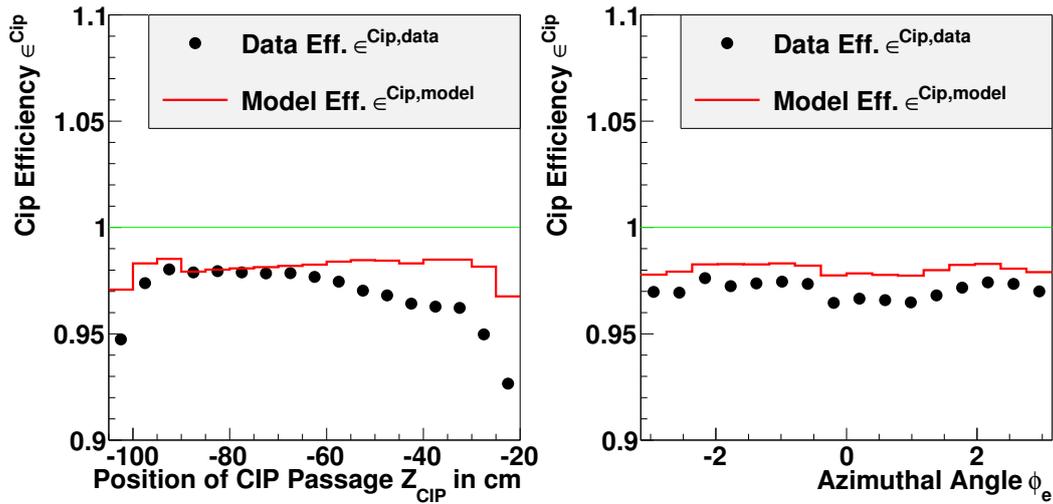
Figure 8.2.3: *CIP efficiency ϵ^{Cip} for two CIP hits along the electron track. Left: As a function of the longitudinal position Z_{CIP} of the electron's CIP passage. Right: As a function of the electron's azimuthal angle ϕ_e.*
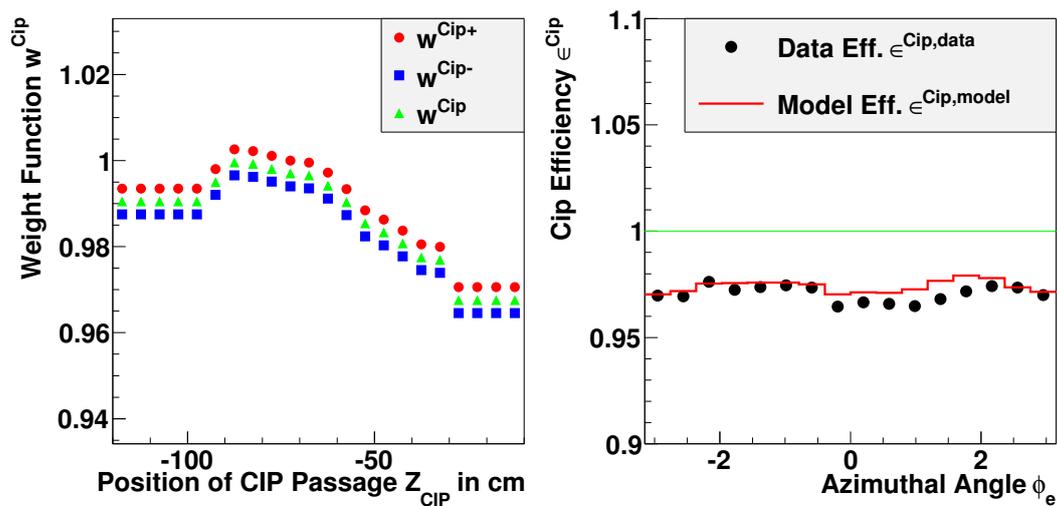
Figure 8.2.4: *Left: CIP efficiency weight functions w^{Cip}, w^{Cip+} and w^{Cip-} as a function of the longitudinal position Z_{CIP} of the electron's CIP passage. Right: CIP efficiency ϵ^{Cip} after reweighting as a function of the electron's azimuthal angle ϕ_e.*
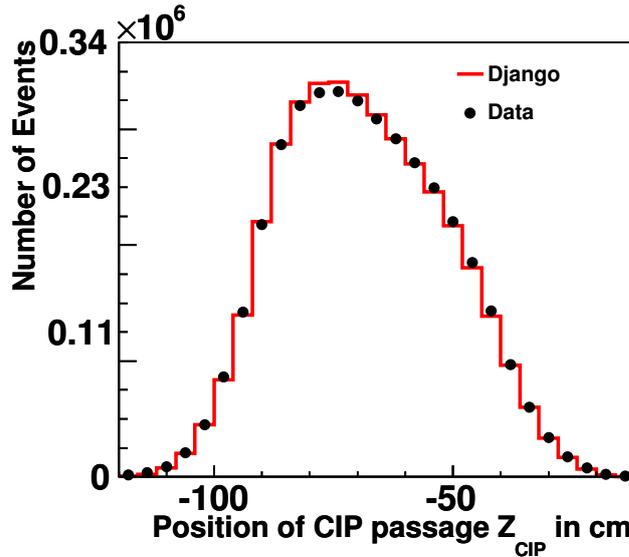
Figure 8.2.5: *Event histogram over the longitudinal position Z_{CIP} of the electron's CIP passage.*

8.2.2   CIP Efficiency

As pointed out in chapter 6, a cut on the number of CIP hits of the electron track has been applied (CIP Hit Cut) in order to suppress background events from neutral particles. Events with $N_{hit} < N_{tr} \equiv 2$ are rejected for small cluster energies ($E_e \lesssim$ 16 GeV). In this subsection, we discuss the CIP efficiency ϵ^{Cip} introduced by this cut.

To define a reference sample, we use only to those events which fullfill the requirement $E_e < E_{tr}^{ref} \equiv 20$ GeV. Since the hadronic background is predominantly observed at small energies $E_e \lesssim 16$ GeV, this sample is regarded as background free. Moreover, we request for the reference sample, that the reconstructed event vertex is of CJC type rather than CIP type, see chapter 6. This is to avoid possible influences of the event selection in the CIP efficiency determination.

In figure 8.2.3 the CIP efficiency is plotted once as function of the azimuthal angle ϕ_e of the SpaCal cluster and once as function of the z-position of the track's passage through the CIP detector (Z_{CIP}). The latter variable is easily calculated from the electron polar angle θ_e and vertex z-position z_{Vtx}:

$$Z_{CIP} = z_{Vtx} + \frac{R_{CIP}}{\tan \theta_e} \tag{8.2.1}$$

Here, $R_{CIP} = 17$ cm denotes the radial position of the inner layer of the CIP detector. Note, that for these computations the conventions of the H1 coordinate system are used.

Note the dropping of the CIP efficiency towards low Z_{CIP} in figure 8.2.3. This reflects the geometry of the CIP detector, which is restricted in its longitudinal range to $z \geq -112$ cm [B$^+$08]. Notice further the drop of effiency at very high values of Z. Also, some systematic variation of the efficiency with ϕ_e is visible. Recall from chapter 3, that the azimuthal structure of the CIP detector shows 8 segments, which differ in performance. This introduces the observed ϕ_e-dependence of ϵ^{Cip}.

Clearly, the data and model efficiencies $\epsilon^{Cip,data}$ and $\epsilon^{Cip,model}$ show some difference, which varies with Z_{CIP}. In particular, this is true for very high and very low values of Z_{CIP}. However, as can be seen from figure 8.2.5, the bulk of the event statistics concentrates on the region $\Omega_{Cip} = [-100$ cm$, -30$ cm$]$. Therefore, the efficiency determination outside this region Ω_{Cip} (region of trust) is not reliable, while its influence on the total error is negligible. Within the region of trust Ω_{Cip}, we observe a difference of up to 2 %, which rises with Z_{CIP}. Note, that almost no large scale ϕ_e-dependance of the difference is observed. The performance of the different sectors of the CIP detector seems very well described by the model. The plot shows an overall difference between $\epsilon^{Cip,data}$ and $\epsilon^{Cip,model}$ of roughly 1 %.

The reweighting $w^{Cip}(Z_{CIP})$ is applied as a function of Z_{CIP}, since the efficiency difference varies prodominantly with Z_{CIP}. For the reweighting procedure, a rather course bin grid $\mathcal{G}_{Cip}$ of 15 bins between -100 cm and 30 cm has been chosen, since the variation of $w^{Cip}(Z_{CIP})$ is small throughout $\mathcal{G}_{Cip}$. For the shifting parameters, we choose $f^{Cip} = 0$ and $p^{Cip} = 0.003$. This means, that a constant offset of 3 permille is applied for the shifts in both directions. This approach can be motivated with figure 8.2.4 (right), in which the *corrected* CIP efficiencies for model ($\tilde{\epsilon}^{Cip,model}$) and data ($\epsilon^{Cip,data}$) is plotted as a function of the electron's azimuthal angle ϕ_e. The overall difference which was visible in figure 8.2.3 is reduced, while fluctuations of ~ 3 permille remain in average. We therefore take this value as a measure for the systematic uncertainty of the model CIP efficiency.

The result is shown in figure 8.2.4 (left), where w^{Cip}, w^{Cip+} and w^{Cip-} are depicted. Note the constant offset of 3 permille bewteen the weight functions. Note further, that for $Z_{CIP} \notin \mathcal{G}_{CipEff}$, the values of the boundary bins of $\mathcal{G}_{CipEff}$ are used.

On analysis level, the CIP efficiency is only applied to low energy events, for which $E_e < E_{tr}^{cut} \equiv 16$ GeV. This is, because the background contamination predominantly occurs in this region of the phase space. Accordingly, the weight functions w^{CipEff}, $w^{CipEff+}$ and $w^{CipEff-}$ are only applied to those low energy events, reducing the systematic uncertainty significantly in the rest of the phase space.

8.2.3 Vertex Efficiency

The determination of the vertex reconstruction efficiency ϵ^{VtxEff} is complicated by the fact, that no clean reference sample can be defined. Recall, that the vertex cut (-35 cm $< z_{Vtx} < 35$ cm) plays a crucial role in the background reduction, particularly by cutting away the satellite bunches. A second problem emerges from the fact, that our event selection exploits two different reconstruction techniques, the CJC vertex reconstruction and the CIP vertex reconstruction. Both have their own efficiency, ϵ^{CjcVtx} and ϵ^{CipVtx}, while our main interest lies in the combined vertex efficiency ϵ^{VtxEff}.

We solve both intricacies at once by the following cross-over approach: We first determine the CJC vertex reconstruction efficiency ϵ^{CjcVtx} using a reference sample comprising all events with CIP vertex. Then, we determine the CIP vertex reconstruction efficiency ϵ^{CipVtx} using a reference sample comprising all events with CJC vertex. Since both reconstruction algorithms use different subdetectors, they are completely independent.

The combined vertex reconstruction efficiency ϵ^{VtxEff} can then be mathematically

derived in the following way:

$$\epsilon^{VtxEff} = \epsilon^{CjcVtx} + \epsilon^{CipVtx} - \epsilon^{CjcVtx} \cdot \epsilon^{CipVtx} \qquad (8.2.2)$$

To obtain an estimate for the statistical uncertainty, we use:

$$\Delta\epsilon^{VtxEff} = \sqrt{\frac{(1 - \epsilon^{VtxEff}) \cdot \epsilon^{VtxEff}}{N^{all}}} \simeq \sqrt{\frac{(1 - \epsilon^{VtxEff}) \cdot \epsilon^{VtxEff}}{N^{VtxEff}}} \qquad (8.2.3)$$

Here, N^{all} denotes the true event count, while N^{VtxEff} refers to all events with a reconstructed vertex either of CIP type or CJC type and with -35 cm $< z_{Vtx} < 35$ cm. Since ϵ^{VtxEff} is very high, $N^{all} \simeq N^{VtxEff}$ is justified.

In figure 8.2.6 (top) the CJC vertex reconstruction efficiency ϵ^{CjcVtx} is plotted, once as function of the SpaCal position R_e of the cluster and once as a function of its energy E_e. Clear discrepancies are visible up to almost 2 % for high R_e and 1% for low E_e. Note, that for events close to the kinematic peak ($E_e \simeq 27.6$ GeV) the CJC vertex efficiency drops rapidly, since this region is dominated by elastic scatterings. In this region, the determination of $\epsilon^{CjcVtx,model}$ and $\epsilon^{CjcVtx,data}$ can not be trusted.

In figure 8.2.6 (center) the CIP vertex reconstruction efficiency ϵ^{CipVtx} is plotted, once as function of the SpaCal position R_e of the cluster and once as a function of its energy E_e. Again, the discrepancies are prominent and reach up to 5 % for low E_e. Note the ascend of ϵ^{CipVtx} with both E_e and R_e.

The combined vertex reconstruction eficiency ϵ^{CipEff} is depiced in figure 8.2.6 (bottom). Notice, that the remaining inefficiency is on a sub-permille scale. Only for small electron energies E_e the inefficiency is almost 1.5 permille. Moreover, the difference between data and model efficiency takes on values up to roughly 1.2 permille only.

Nevertheless, a reweighting of the vertex reconstruction efficiency has been done. Since only the E_e-dependence of the difference $\epsilon_{data}^{VtxEff} - \epsilon_{model}^{VtxEff}$ is potentially significant, the weights $w^{VtxEff}(E_e)$ are defined as a function of E_e. For the reweighting, the region $\Omega_{VtxEff} = [11$ GeV$, 27$ GeV$]$ (region of trust) has been devided in 17 bins ($\mathcal{G}_{VtxEff}$). For the shift parameters, we set $f^{VtxEff} = 0.5$ and $p^{VtxEff} = 0$ as usual. The result is shown in figure 8.2.2, where w^{VtxEff}, $w^{VtxEff+}$ and $w^{VtxEff-}$are depicted for the bins of $\mathcal{G}_{VtxEff}$.
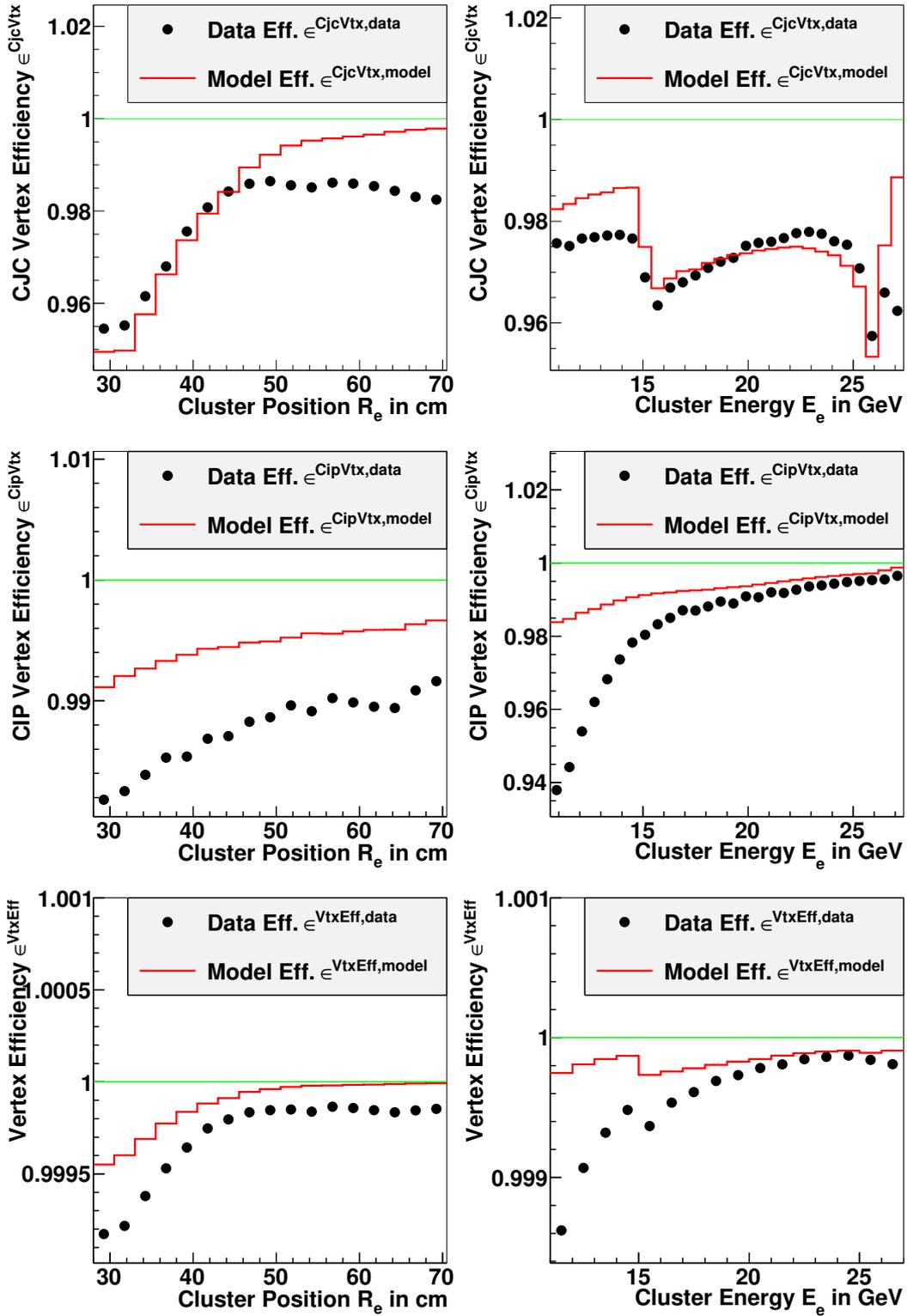
Figure 8.2.6: *Vertex reconstruction efficiency. Top: The CJC vertex reconstruction efficiency ϵ^{CjcVtx}. Center: The CIP vertex reconstruction efficiency ϵ^{CipVtx}. Bottom: The combined vertex reconstruction efficiency ϵ^{CjcVtx}. The left column shows the efficiencies as a function of the radial position R_e of the SpaCal electron, the right column as a function of its electromagnetic energy E_e.*
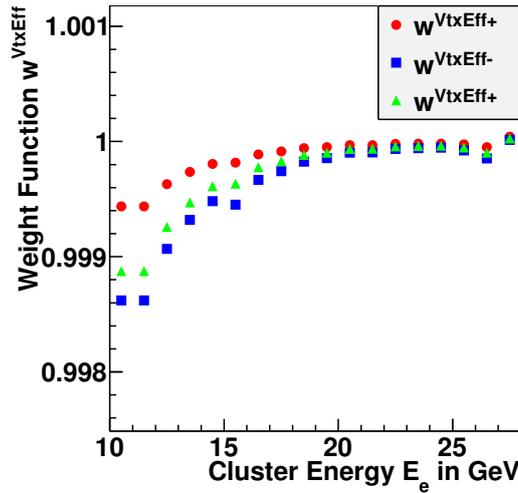
Figure 8.2.7: *Vertex reconstruction efficiency weight functions w^{VtxEff}, $w^{VtxEff+}$ and $w^{VtxEff-}$ are depicted as a function of the electromagnetic energy E_e of the SpaCal electron.*

8.2.4 ECRA Cut Efficiency

As pointed out in chapter 6, a cut on the SpaCal cluster size has been applied (ECRA Cut) in order to suppress background from hadronic particles. Events with $ECRA_e > ECRA_{tr}(R_e, E_e)$ are rejected, while $ECRA_{tr}(R_e, E_e)$ is a function on the radial position R_e of the SpaCal cluster and its electromagnetic energy E_e. In this subsection, we discuss the ECRA efficiency ϵ^{Ecra} introduced by this cut. Recall, that the variable $ECRA_e$ is not well described by the Monte Carlo simulation (see figure 7.4.3), such that an efficiency difference is expected.

In order to construct a clean reference sample, we search specifically for QED Compton events. These events have a distinct signature, exhibiting two clusters which are back-to-back in azimuthal angle ϕ and fulfill $(E - p_z)_1 + (E - p_z)_2 = 2E_{beam}$, see chapter 6. We therefore search for pairs of SpaCal clusters, which fulfill the following conditions:

$$-10° < \phi_1 - \phi_2 + 180° < 10° \tag{8.2.4}$$

$$-10 \text{ GeV} < (E - p_z)_1 + (E - p_z)_2 - 2 \cdot E_{beam} < 10 \text{ GeV} \tag{8.2.5}$$

$$7 \text{ GeV} < E_{1,2} < 22 \text{ GeV} \tag{8.2.6}$$

Recall from chapter 6, that we also have $E_1 > 11$ GeV. The above selection is expected to significantly reduce the background contamination. Of course, the QEDC Cut from chapter 6 is relaxed for this sample.

In order to enhance the statistics of the reference sample, the data sample is extended to the electron run periods in 2005 and 2006. The complete run range of the sample is 399101 to 500611. Here we use, that the ECRA cut efficiency is expected to be time invariant. The same good run selection is used, which has been described in 6. For the integrated luminosity, we obtain $\mathcal{L}_{data}^{QEDC} \simeq 301.5$ pb^{-1}. For the Monte Carlo reference sample, a `DJANGOH` generation is used, which comprises QED Compton events only[1]. The integerated luminosity is $\mathcal{L}_{model}^{QEDC} \simeq 1124.8$ pb^{-1}.

---

[1]Requests 7117, 7119.

In figure 8.2.8 the ECRA cut efficiency ϵ^{Ecra} is plotted, once as function of the SpaCal position R_e of the cluster and once as a function of its energy E_e. Note, that inefficiencies only appear in the area of small E_e and small R_e, since at higher values the ECRA cut is relaxed. As expected, discrepancies between ϵ_{data}^{Ecra} and ϵ_{model}^{Ecra} are clearly visible, which take on values up to ~ 1.5 % for low E_e.

Since the discrepancies depend on both R_e and E_e, a two dimensional weight function $w^{Vtx}(R_e, E_e)$ needs to be introduced. To do that, a two dimensional bin grid $\mathcal{G}_{Ecra}$ is defined, comprising 3 bins in R_e in 4 in E_e. The bin boundaries in the direction of R_e and E_e are:

$$R_e^{bnd} = 28 \text{ cm, } 40 \text{ cm, } 55 \text{ cm, } 72 \text{ cm} \tag{8.2.7}$$

$$E_e^{bnd} = 11 \text{ GeV, } 12.5 \text{ GeV, } 14 \text{ GeV, } 16 \text{ GeV} \tag{8.2.8}$$

The rather coarse choice of the binning accounts for the limited statistics of the QED Compton sample. The bin boundaries are chosen to reflect the structure of the threshold function $ECRA_{tr}(R_e, E_e)$. For the systematic uncertainties, we set $f^{Ecra} = 0.5$ and $p^{Ecra} = 0$. The result is shown in figure 8.2.9, where w^{Ecra}, w^{Ecra+} and w^{Ecra-} are depicted for the bins of $\mathcal{G}_{Ecra}$.
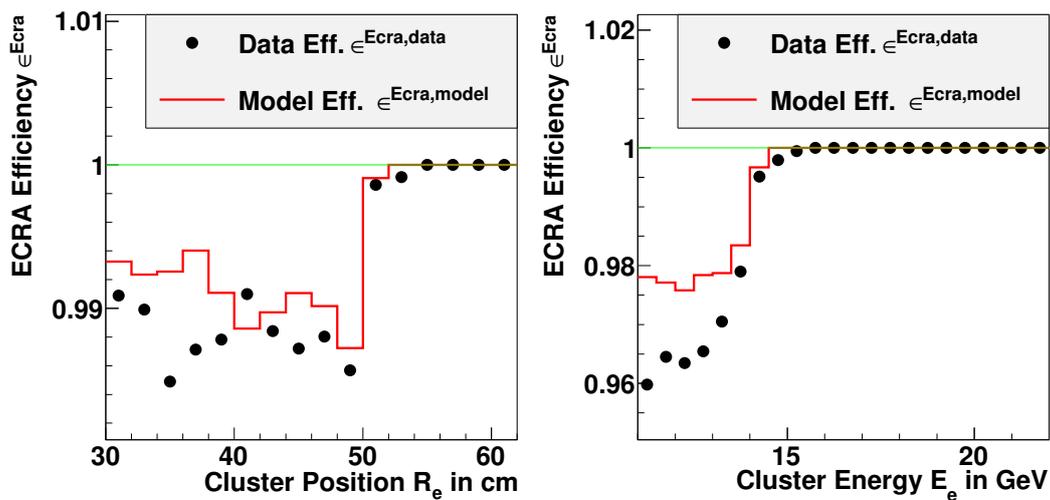
Figure 8.2.8: *ECRA cut efficiency ϵ^{Ecra}. Left: As a function of the radial position R_e of the SpaCal electron. Right: As a function of the electromagnetic energy E_e of the SpaCal electron.*
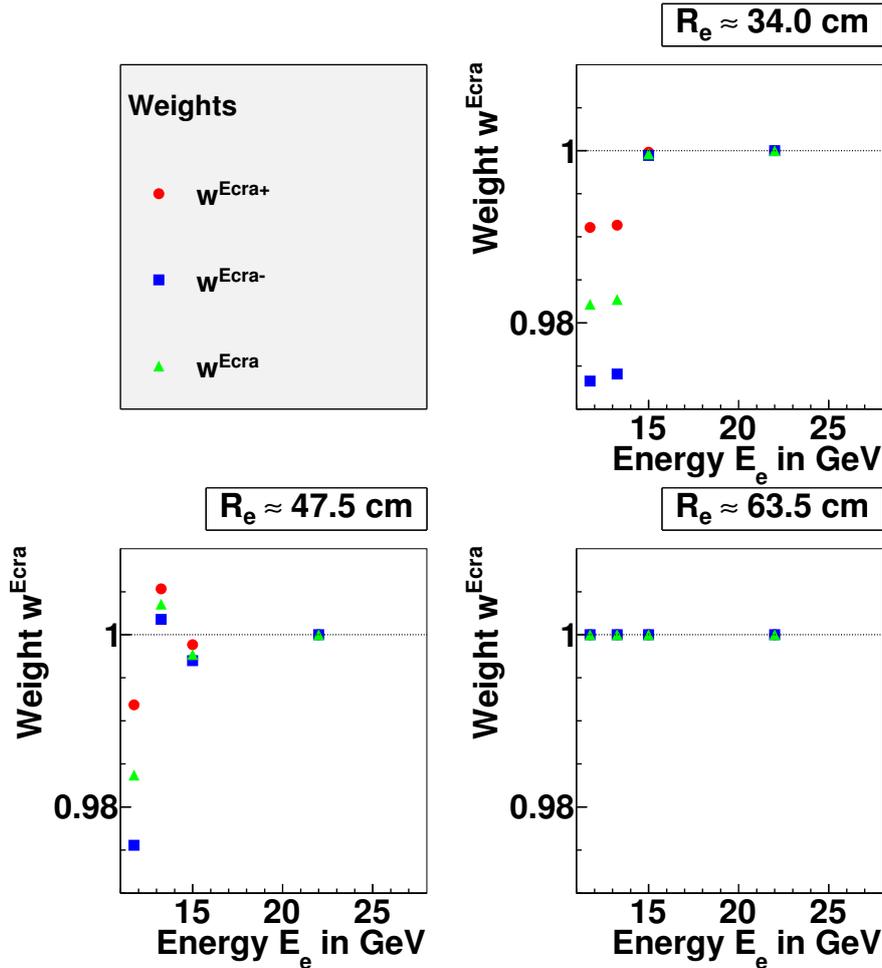
Figure 8.2.9: *ECRA cut efficiency weight functions w^{Ecra}, w^{Ecra+} and w^{Ecra-} are depicted as a function of the radial position R_e of the SpaCal electron and its electromagnetic energy E_e.*

8.3 Z Vertex Reweighting

As pointed out in chapter 7, the Monte Carlo simulation fails to match the Z vertex shape distribution, see figure 8.3.1 (left). This is remedied by introducing a shape weight function $w^{VtxRew}(z_{Vtx})$ that depends on the longitudinal position z_{Vtx} of the primary vertex. The result can be seen in figure 7.4.2 (center right), which features a good agreement of the distributions for data and Monte Carlo. The reweighting is done very similar to the efficiency correction.

8.3.1 The Z vertex weight function

The weight function $w^{VtxRew}(z_{Vtx})$ is directly obtained as a ratio of the data distribution $N^{data}(z_{Vtx})$ and the Monte Carlo distribution $N^{model}(z_{Vtx})$. Note that, unlike for efficiency corrections, a pure *shape reweighting* is applied. This means, both numerator

and denominator have to be normalized to their luminosity. We therefore define:

$$w^{VtxRew}\left(z_{Vtx}\right) = \left(\frac{\sum N^{model}\left(z_{Vtx}\right)}{\sum N^{data}\left(z_{Vtx}\right)}\right) \cdot \frac{N^{data}\left(z_{Vtx}\right)}{N^{model}\left(z_{Vtx}\right)} \qquad (8.3.1)$$

Here, the sum runs over all z_{Vtx}-bins in the bin grid $\mathcal{G}_{VtxRew}$ to be defined. Note, that the weight function $w^{VtxRew}\left(z_{Vtx}\right)$ can take on values both greater and less than 1.

The application of $w^{VtxRew}\left(z_{Vtx}\right)$ to the Monte Carlo simulation is done in the way described in subsection 8.1.1. Note however, that the Z Vertex shape reweighting must also be applied to the generator level distribution $\vec{N}^{gen}$, such that we have:

$$\vec{N}^{gen} \rightarrow \tilde{\vec{N}}^{gen}$$

This is critical to the preservation of the correct reconstruction efficiencies in the unfolding process.

From 8.3.1 (left) it is clear, that for z_{Vtx} with large absolute value $|z_{Vtx}|$, the event statistic is poor. Large systematic uncertainties on the weight functions have to be expected. Therefore, a restriction of the reweighting prescription to the luminous region of the detector should be applied. We do this formally by introducing a *region of trust* $\Omega_{VtxRew} \equiv [z_{Vtx}^{min}, z_{Vtx}^{max}]$ and refine the definition (8.3.1) in the following way ($\alpha \in \{\text{model, data}\}$):

$$\tilde{N}^{\alpha}\left(z_{Vtx}\right) = \begin{cases} N^{\alpha}\left(z_{Vtx}^{min}\right) & \forall z_{Vtx} < z_{Vtx}^{min} \\ N^{\alpha}\left(z_{Vtx}\right) & \forall z_{Vtx} \in \Omega_{Rew} \\ N^{\alpha}\left(z_{Vtx}^{max}\right) & \forall z_{Vtx} > z_{Vtx}^{max} \end{cases} \qquad (8.3.2)$$

Then, let $\tilde{w}^{VtxRew}\left(z_{Vtx}\right)$ be given analog to (8.3.1). This essentially means, the weight function is constrained to the region Ω_{VtxRew}, while for events with $z_{Vtx} \notin \Omega_{VtxRew}$ the weights from the boundaries of Ω_{VtxRew} are used. From now on, with the symbol $w^{VtxRew}\left(z_{Vtx}\right)$, we *implicitly* understand a weight function according to (8.3.2) and (8.3.1).
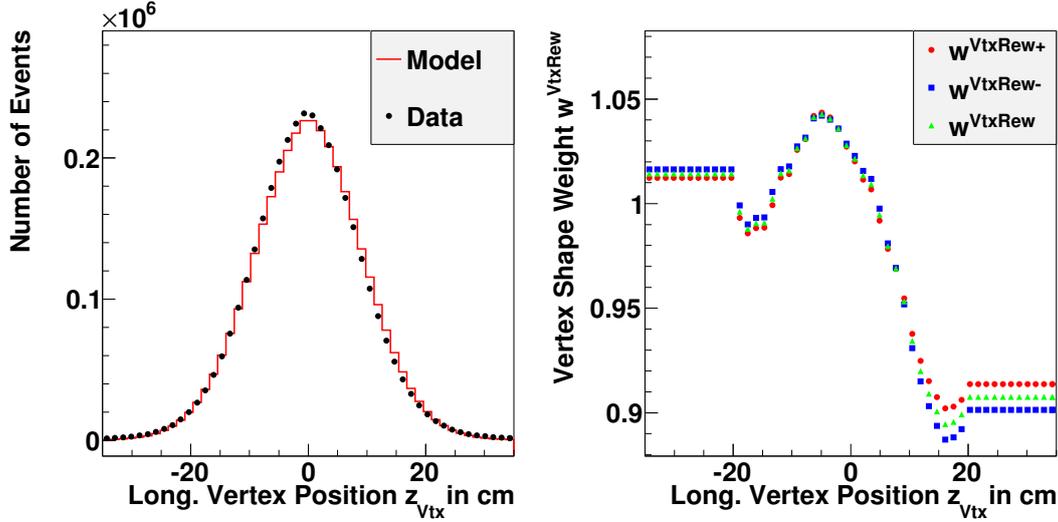
Figure 8.3.1: *Left: Comparison of the distributions of the l z-Vertex distribution for data and Monte Carlo before reweighting, $N^{data}(z_{vtx})$ and $N^{model}(z_{vtx})$. The model distribution is normalized to the data distribution according to (7.1.6). Right: Weights w^{VtxRew}, $w^{VtxRew+}$ and $w^{VtxRew-}$ for the z-Vertex shape reweighting. For the z Vertex distribution after reweighting, see figure 7.4.2 (center right).*

8.3.2   Systematic Uncertainty of Z Vertex Reweighting

In order to account for the systematic uncertainty of z Vertex reweighting, *shifted weight functions* $w^{VtxRew+}(z_{Vtx})$ and $w^{VtxRew-}(z_{Vtx})$ are introduced. Completely analog to subsection 8.1.2, two shifted response matrices $\boldsymbol{N}^{VtxRew\pm}$ and two shifted detector level distributions $\vec{N}^{rec,VtxRew\pm}$ are obtained. Moreover, we obtain two shifted generator level distributions $\vec{N}^{gen,VtxRew\pm}$, which are important to preserve the reconstruction efficiency in the unfolding process.

Completely analog to subsection 8.1.2, the shifted weight functions $w^{VtxRew\pm}(z_{Vtx})$ are constructed by shifting the model z-Vertex distribution $N^{model}(z_{Vtx})$ up and down, such that we obtain $N^+(z_{Vtx})$ and $N^-(z_{Vtx})$. These shifts in turn propagate to the weight functions $w^{VtxRew+}(z_{Vtx})$ and $w^{VtxRew-}(z_{Vtx})$ by means of (8.3.1). We set for the shifted model distributions:

$$N^{\pm}(z_{Vtx}) = N^{data}(z_{Vtx}) \pm$$

$$f^{VtxRew} \cdot \left(N^{data}(z_{Vtx}) \cdot \left(\frac{\sum N^{model}(z_{Vtx})}{\sum N^{data}(z_{Vtx})} \right) - N^{model}(z_{Vtx}) \right) \pm p^{VtxRew} \qquad (8.3.3)$$

Note the normalization factor in (8.3.3), the sums of which run over all z_{Vtx}-bins in the grid $\mathcal{G}_{VtxRew}$. As always, the functional values outside the luminous region of the detector, i.e. for $z_{Vtx} \notin [z_{Vtx}^{min}, z_{Vtx}^{max}]$, will be fixed to the boundary values:

$$\tilde{N}^{\pm}(z_{Vtx}) = \begin{cases} N^{\pm}(z_{Vtx}^{min}) & \forall z_{Vtx} < z_{Vtx}^{min} \\ N^{\pm}(z_{Vtx}) & \forall z_{Vtx} \in \Omega_{Rew} \\ N^{\pm}(z_{Vtx}^{max}) & \forall z_{Vtx} > z_{Vtx}^{max} \end{cases} \qquad (8.3.4)$$

With $\tilde{N}^{\pm}(z_{Vtx})$ from (8.3.3) and $\tilde{N}^{data}(z_{Vtx})$ from (8.3.2) the shifted weight functions $w^{VtxRew\pm}(z_{Vtx})$ can be calculated according to (8.3.1). With this definition, the normalization of the Z vertex shape is preserved even when the shifted weight functions are applied.

8.3.3   Estimation of Z Vertex Weights

In order to estimate the Z vertex weights, the data and Monte Carlo distributions are directly compared, see figure 8.3.1 (left). A slight horizontal shift between the distributions is visible.

The weights w^{VtxRew} are determined for the region $\Omega_{VtxRew} = [\text{-20 cm}, +20 \text{ cm}]$, while a bin grid $\mathcal{G}_{VtxRew}$ with 30 bins has been chosen. For the shifting parameters, we set $f_{VtxRew} = 0.1$ and $p_{VtxRew} = 0$, reflecting the abundant statistics of our data sample. The result is shown in figure 8.3.1 (right), where w^{VtxRew}, $w^{VtxRew+}$ and $w^{VtxRew-}$ are depicted for the bins of $\mathcal{G}_{Rew}$. Note, that the correction can take on values up to 9%.

Chapter 9

Calibration and Alignment Tests

In Chapter 3 the general SpaCal calibration procedure was covered. Here, calibration tests are performed with regard to the electron variables E_e and θ_e. We particularly study, how both variables are reproduced in the Monte Carlo simulation. The discussion spans both the absolute calibration scale and the detector resolution. From that, measures for the systematic uncertainties are derived, that account for the detector calibration and resolution.

9.1 Electron Energy Scale and Resolution

In order to study the calibration of the electron energy E_e, we compare this quantity with the energy obtained from the double angle method E_{DA}, see equation (3.5.7). Note, that both quantities do not necessarily match, since the SpaCal cluster energy can be lowered due to photon radiation off the electron line.

In figure 9.2.1, the shift

$$\delta E := E_e - E_{DA} \tag{9.1.1}$$

has been plotted for different radial SpaCal regions. For the binning in R_e, 10 bins with the width of 4 cm have been chosen, ranging from 30 cm to 70 cm. The data sample for this calibration study is restricted to $E_{DA} > 25$ GeV to reduce the influence of radiation effects.

As can be seen from figure 9.2.1, both the data and Monte Carlo distributions do not peak at zero. Their maxima are rather shifted by 200 to 500 MeV towards negative values, depending on the R_e-bin. Moreover, the distributions are not fully symmetric, but exhibit tails reflecting the residual influence of radiative effects. For each bin i, the mean value $\mu^i\left(\delta E\right)$ is shifted by roughly 1 GeV and its root mean square $\delta^i\left(\delta E\right)$ takes on values around 1 GeV as well. Note however, that all shifts, resolution widths and tails of the distributions are well described by the Monte Carlo simulation. One obtains:

$$\Delta\mu^i \equiv \mu^i_{data}\left(\delta E\right) - \mu^i_{data}\left(\delta E\right) \simeq 15 \text{ - } 100 \text{ MeV} \tag{9.1.2}$$

$$\Delta\delta^i \equiv \delta^i_{data}\left(\delta E\right) - \delta^i_{model}\left(\delta E\right) \simeq 10 \text{ - } 60 \text{ MeV} \tag{9.1.3}$$

The estimation of the systematic uncertainties on the unfolding result $\vec{x}$ is done according to section 4.8. Two shifted matrices $\boldsymbol{N}^{E_e\pm}$ are produced by shifting the electron energy scale up and down for each event. By the above discussion, the following

constant offset is motivated for the input shift:

$$(\Delta E_e)_\pm = \pm\ 50 \text{ MeV} \tag{9.1.4}$$

For the electron energy resolution $Res(E_e)$, a single shifted matrix $\boldsymbol{N}^{Res(E_e)}$ is produced by shifting the resolution up by:

$$\Delta Res(E_e) = 40 \text{ MeV} \tag{9.1.5}$$

The effect of those shifts on the unfolding result is described in chapter 10.

9.2   Electron Polar Angle Calibration and Resolution

We now test the calibration of the polar angle θ_e of the electron clusters, which is essentially an alignment check for the SpaCal. We compare θ_e to the polar angle θ_{Track} from the electron track. The angle θ_{Track} is particularly reliable for tracks which show hits in the COZ subdetector, because the alignment of both the CJC and the SpaCal detector is done relative to the COZ position. Electron clusters in the outer region of the SpaCal can often be linked to a track with COZ hits.

In figure 9.2.2, the shift

$$\delta\theta := \theta_e - \theta_{Track} \tag{9.2.1}$$

has been plotted for different energy regions. For the binning in E_e, 8 bins with the width of 2 GeV have been chosen, ranging from 11 GeV to 27 GeV. The data sample for this alignment study is restricted to events, for which a vertex fitted track can be matched to the electron cluster. Moreover, the electron track is required to have at least 2 hits in the COZ detector.

As can be seen from figure 9.2.2, the root mean square of the distribution typically takes on values between 3 and 4 mrad. The consistency of data and model distributions is acceptable and one obtains:

$$\Delta\mu_i \equiv \mu_{data}^i\left(\delta\theta\right) - \mu_{data}^i\left(\delta\theta\right) \simeq 0.7\ \text{-}\ 1.1 \text{ mrad} \tag{9.2.2}$$

$$\Delta\delta_i \equiv \delta_{data}^i\left(\delta\theta\right) - \delta_{data}^i\left(\delta\theta\right) \simeq 0.3\ \text{-}\ 0.5 \text{ mrad} \tag{9.2.3}$$

Again, the estimation of the systematic uncertainties on the unfolding result $\vec{x}$ is done according to section 4.8. Two shifted matrices $\boldsymbol{N}^{\theta_e\pm}$ are produced by shifting the electron polar angle scale up and down for each event. For the shift, a constant offset of

$$\Delta\theta_e = \pm 1 \text{ mrad} \tag{9.2.4}$$

is chosen. For the electron polar angle resolution $Res(\theta_e)$, a single shifted matrix $\boldsymbol{N}^{Res(\theta_e)}$ is produced by shifting the resolution up by:

$$\Delta Res(\theta_e) = 0.4 \text{ mrad} \tag{9.2.5}$$

The result of this procedure is described in chapter 10.
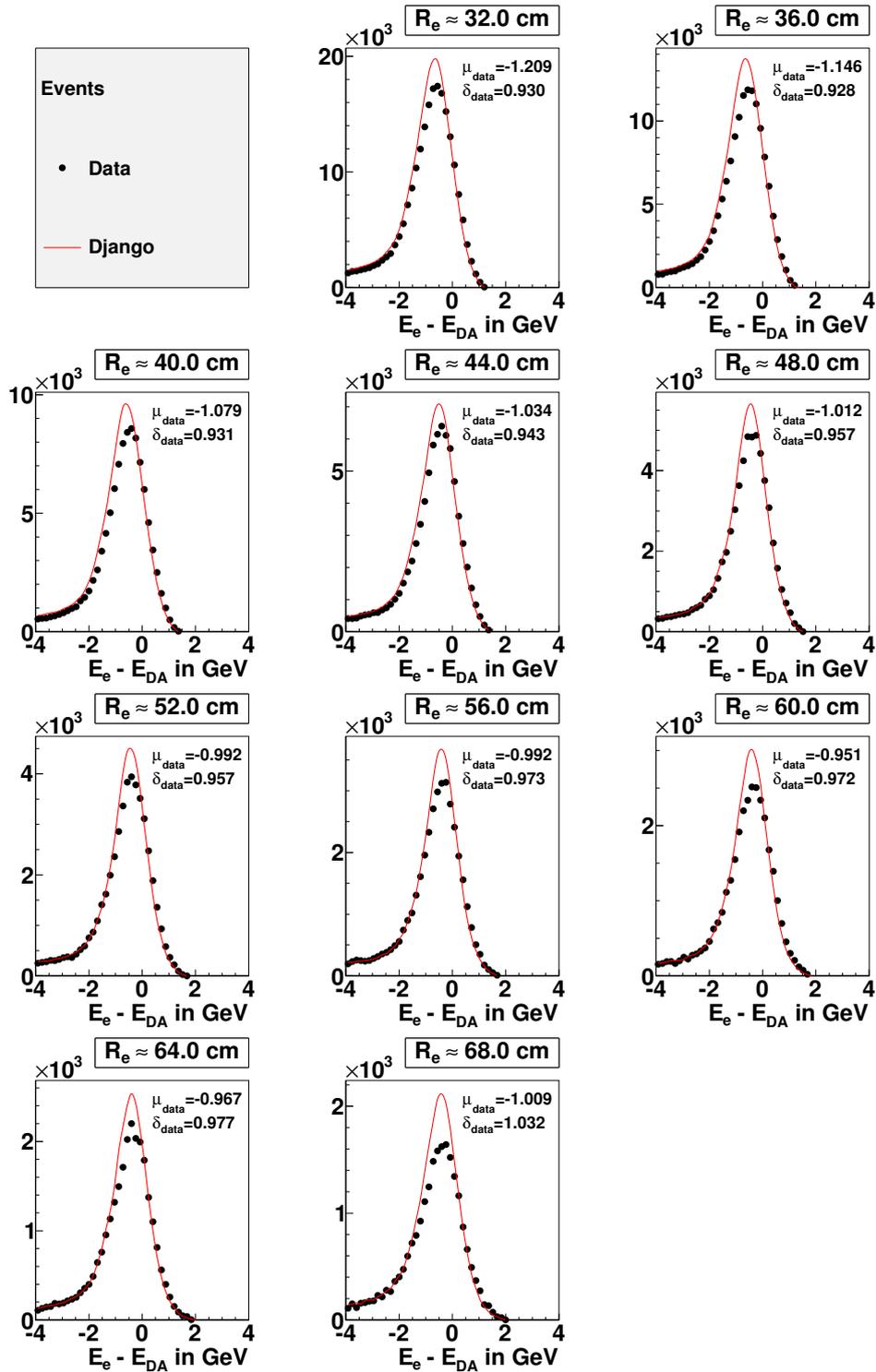
Figure 9.2.1: *Comparison of SpaCal cluster energy E_e and double angle energy E_{DA}. The distribution of the shift $\delta E := E_e - E_{DA}$ for different bins in R_e is shown. For this event sample, $E_e > 25$ GeV is required. For each bin i, also the mean values μ_i and the root mean squares δ_i are given for the data distribution.*
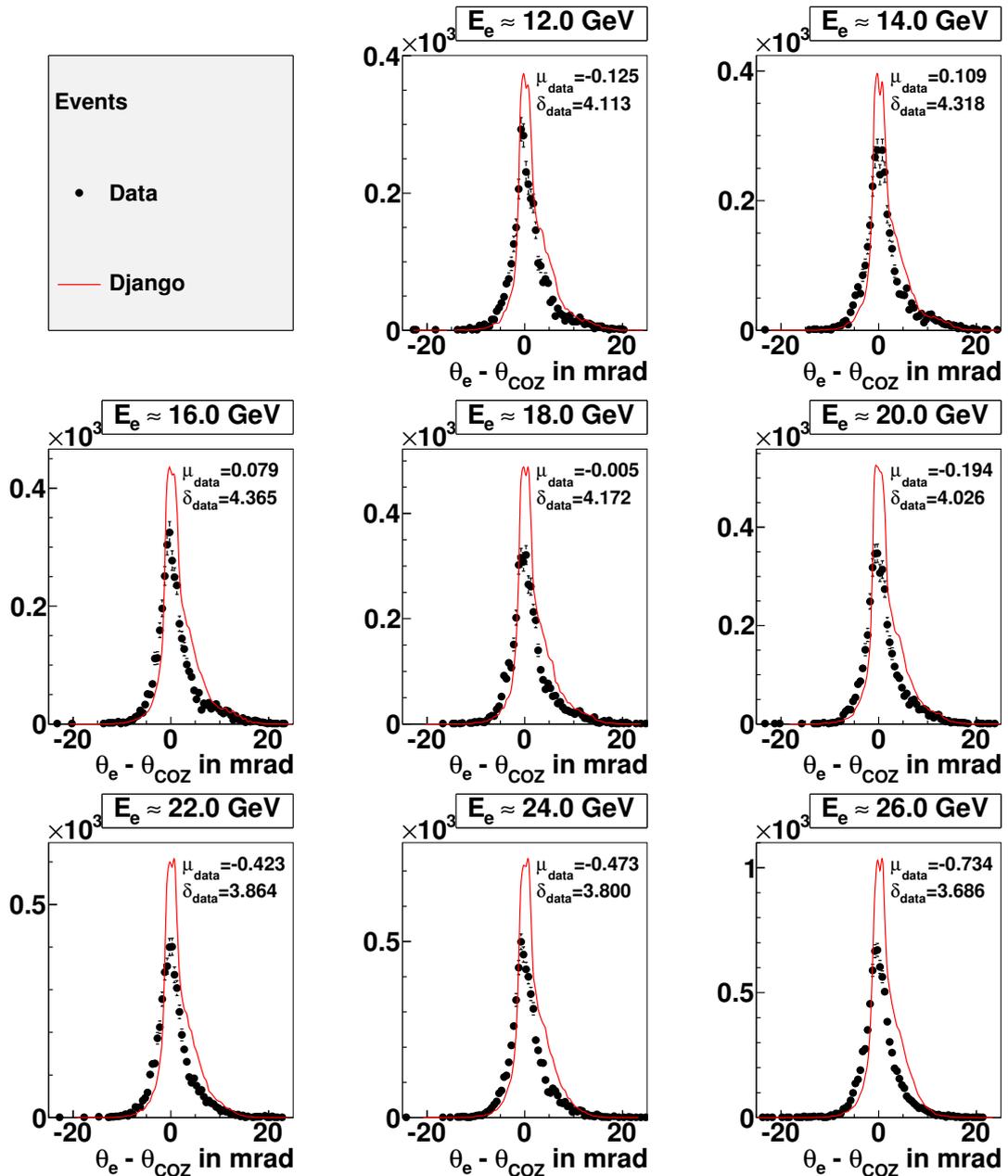
Figure 9.2.2: *Comparison of SpaCal cluster polar angle θ_e and the track angle θ_{Track}. The distribution of the shift $\delta\theta := \theta_e - \theta_{Track}$ for different bins in E_e is shown. For this event sample, vertex fitted tracks with at least 2 COZ hits are required, which have to match the electron cluster. For each bin i, also the mean values μ_i and the root mean squares δ_i are given for the data distribution.*

Chapter 10

Unfolding the data

In this chapter we apply the unfolding procedure introduced in chapter 4 to the measurement of the double differential neutral current $ep \rightarrow eX$ cross section in the backward region of the H1 detector.

We start with the introduction of the generator and detector level bin grids $\mathcal{G}_{gen}$ and $\mathcal{G}_{det}$, we study both efficiency and migration for these binnings and introduce the response matrix.

To determine the level of regularization, the L curve analysis (subsection 4.4.1), the algebraic analysis (subsection 4.4.2) and the global correlation method (subsection 4.7.2) are performed and compared. We will argue, that the result of the algebraic analysis according to V. BLOBEL provides the appropriate choice of the regularization parameter τ.

We will study the unfolding result with emphasis on the differences to the standard bin-by-bin method, both concerning the cross section itself and its statistical uncertainty. Moreover, we will study the systematic shifts on the result that are introduced by the different systematic error sources.

10.1   Bin Grid and Response Matrix

The generator bin grid $\mathcal{G}_{gen}$ chosen for this analysis features 18 bins in the direction of Q_e^2 and 10 bins in the direction of y_e. Together with its overflow bins, $\mathcal{G}_{gen}$ comprises 240 bins $\mathcal{B}_{i_1,i_2}$ in total. Each bin $\mathcal{B}_{i_1,i_2}$ is given a unique bin number $i = \nu(i_1, i_2)$, which is defined in (4.3.7). For better orientation, the bin numbering scheme is depicted in figure 10.1.1. In Table 10.1, the bin boundaries are given.

From figure 10.1.2 it can be seen, that the efficiency varies strongly throughout the bin grid. This is due to the boundaries of the detector acceptance, which lead to "blind spots" in some phase space regions. According to the arguments in subsection 4.5.3, we introduce an efficiency threshold of $\epsilon_{tr} = 70\%$. Figure 10.1.1 shows the bins with sufficient efficiency in green (*quotable bins*), while low efficiency bins are shown in yellow and overflow bins are supressed. In total, we have 106 quotable bins, 60 overflow bins and 74 low efficiency bins in the generator level bin grid $\mathcal{G}_{gen}$. Note, that we have no bins with vanishing efficiency (*zero efficiency bins)*. Table 10.3 features a summary of the different bin types.

In this analysis, the low efficiency bins are *not* excluded from unfolding ("cut-last-approach", compare subsection 4.5.5). Therefore, all 240 generator level bins will be

used for the unfolding process. However, low efficiency bins and overflow bins will be discarded later.

In Figure 10.1.5 the different efficiencies defined in chapter 4 are drawn and compared. Note, that for some bins with low y the bin grid efficiency takes on values greater than one.

Figure 10.1.2 shows the purity p_{i_1,i_2} and the stability s_{i_1,i_2} for each bin $\mathcal{B}_{i_1,i_2}$. Note, that inefficient bins and overflow bins are suppressed. Both stability and purity take on values around 60% for all quotable bins. Moreover, it can be seen, that both stability and purity are constant troughout the efficient region of the bin grid. Therefore, the condition (4.3.14) of regional conformity is fulfilled. From table 10.1 the one dimensional migration effects can be obtained. Both purity and stability are of comparable size for both directions, such that the condition of directional conformity (4.3.15) can be regarded as fulfilled. We conclude, that the application of curvature regularization as defined in (4.3.10) is justified.

For the detector level bin grid $\mathcal{G}_{det}$ a finer binning must be chosen, which we achieve by dividing each bin in half for each direction, see Table 10.2. Therefore, we end up with 36 bins in the direction of Q^2 and 20 bins in the direction of y, giving a total number of 836 bins $\mathcal{B}^{det}_{j_1,j_2}$ including the overflow bins. Among those are 310 *empty bins.* In particular, the 116 overflow bins are empty, which is due to the phase space cuts (6.3.4) and (6.3.3) applied on analysis level. See also table 10.3 for a summary of the different event types.

Figure 10.1.3 shows the detector level distributions for the measured data $\vec{N}^{raw}$ and the Monte Carlo model $\vec{N}^{rec}$ as a function of the bin number. The steeply falling behaviour with both Q_e^2 and y_e can be directly observed. The Monte Carlo and distribution has been normalized to the data luminosity according to (7.1.4).

In Figure 10.1.4 the scaled unfolding matrix $\boldsymbol{N}^{MC}$ is depicted. The migration effects are clearly visible. Note, that the lower right corner is less populated than the upper left region of the plot. This shows, that migration predominantly occurs from lower to higher values of y_e and Q_e^2. This is expected due to the steeply falling behaviour of the double differential cross section.

10.2   Adjustment of the Regularization

In the context of this analysis, curvature regularization has been applied. This is done using definition (4.3.25) for two dimensions. In Figure 10.1.6, the matrix $\boldsymbol{L}^2$ is shown. We immediatly recognize the structure of (4.3.11).

10.2.1   L Curve Analysis

The outcome of a straightforward L curve analysis according to subsection 4.4.1 is shown in 10.2.1. This L curve analysis is done with $N_{scan} = 500$ scan points, yielding:

$$\log \tau_L \simeq -3.46 \tag{10.2.1}$$

Note, that for this value of τ, we obtain $\chi_A^2 \simeq 497$.

Note, that $m = 526$ detector level bins and $n = 240$ generator level bins enter the unfolding process. For the χ^2-minimization process, this leaves

$$n_{DOF} = m - n = 286 \tag{10.2.2}$$

Q_e^2-Bin	Left	Right	Stability s^{Q^2}	Purity p^{Q^2}	Status
0	$-\infty$	10.0 GeV^2	0.000	0.000	OF-Bin
1	10.0 GeV^2	12.5 GeV^2	0.750	0.779	
2	12.5 GeV^2	14.3 GeV^2	0.735	0.729	
3	14.3 GeV^2	16.0 GeV^2	0.703	0.695	
4	16.0 GeV^2	18.5 GeV^2	0.766	0.755	
5	18.5 GeV^2	21.0 GeV^2	0.742	0.733	
6	21.0 GeV^2	23.5 GeV^2	0.722	0.715	
7	23.5 GeV^2	26.5 GeV^2	0.777	0.739	
8	26.5 GeV^2	30.0 GeV^2	0.757	0.750	
9	30.0 GeV^2	34.0 GeV^2	0.756	0.758	
10	34.0 GeV^2	38.0 GeV^2	0.734	0.738	
11	38.0 GeV^2	43.0 GeV^2	0.759	0.763	
12	43.0 GeV^2	49.0 GeV^2	0.774	0.779	
13	49.0 GeV^2	55.0 GeV^2	0.756	0.760	
14	55.0 GeV^2	62.0 GeV^2	0.765	0.774	
15	62.0 GeV^2	71.0 GeV^2	0.790	0.799	
16	71.0 GeV^2	88.0 GeV^2	0.856	0.868	
17	88.0 GeV^2	110.0 GeV^2	0.849	0.873	
18	110.0 GeV^2	133.4 GeV^2	0.702	0.901	
19	133.4 GeV^2	$+\infty$	0.000	0.000	OF-Bin

y_e-Bin	Left	Right	Stab. s^y	Pur. p^y	Status
0	$-\infty$	0.06	0.376	0.756	OF-Bin
1	0.06	0.12	0.765	0.661	
2	0.12	0.18	0.722	0.699	
3	0.18	0.24	0.731	0.714	
4	0.24	0.30	0.743	0.724	
5	0.30	0.35	0.710	0.691	
6	0.35	0.40	0.720	0.700	
7	0.40	0.45	0.733	0.710	
8	0.45	0.50	0.749	0.721	
9	0.50	0.55	0.756	0.738	
10	0.55	0.62	0.921	0.816	
11	0.62	$+\infty$	0.000	0.000	OF-Bin

Table 10.1: *Generator level binning in Q_e^2 (top) and y_e (bottom). Moreover, the one dimensional purities and stabilities are listed.*

Q_e^2-Bin	Left	Right	Status
0	$-\infty$	10.0 GeV^2	OF-Bin
1	10.0 GeV^2	11.2 GeV^2	
2	11.2 GeV^2	12.5 GeV^2	
3	12.5 GeV^2	13.4 GeV^2	
4	13.4 GeV^2	14.3 GeV^2	
5	14.3 GeV^2	15.1 GeV^2	
6	15.1 GeV^2	16.0 GeV^2	
7	16.0 GeV^2	17.2 GeV^2	
8	17.2 GeV^2	18.5 GeV^2	
9	18.5 GeV^2	19.8 GeV^2	
10	19.8 GeV^2	21.0 GeV^2	
11	21.0 GeV^2	22.2 GeV^2	
12	22.2 GeV^2	23.5 GeV^2	
13	23.5 GeV^2	24.0 GeV^2	
14	24.0 GeV^2	26.5 GeV^2	
15	26.5 GeV^2	28.2 GeV^2	
16	28.2 GeV^2	30.0 GeV^2	
17	30.0 GeV^2	32.0 GeV^2	
18	32.0 GeV^2	34.0 GeV^2	
19	34.0 GeV^2	36.0 GeV^2	
20	36.0 GeV^2	38.0 GeV^2	
21	38.0 GeV^2	40.5 GeV^2	
22	40.5 GeV^2	43.0 GeV^2	
23	43.0 GeV^2	46.0 GeV^2	
24	46.0 GeV^2	49.0 GeV^2	
25	49.0 GeV^2	52.0 GeV^2	
26	52.0 GeV^2	55.0 GeV^2	
27	55.0 GeV^2	58.5 GeV^2	
28	58.5 GeV^2	62.0 GeV^2	
29	62.0 GeV^2	66.5 GeV^2	
30	66.5 GeV^2	72.0 GeV^2	
31	72.0 GeV^2	79.5 GeV^2	
32	79.5 GeV^2	88.0 GeV^2	
33	88.0 GeV^2	99.0 GeV^2	
34	99.0 GeV^2	110.0 GeV^2	
35	110.0 GeV^2	120.0 GeV^2	
36	120.0 GeV^2	133.4 GeV^2	
37	133.4 GeV^2	$+\infty$	OF-Bin

y_e-Bin	Left	Right	Status
0	$-\infty$	0.06	OF-Bin
1	0.06	0.09	
2	0.09	0.12	
3	0.12	0.15	
4	0.15	0.18	
5	0.18	0.21	
6	0.21	0.24	
7	0.24	0.27	
8	0.27	0.30	
9	0.30	0.32	
10	0.32	0.35	
11	0.35	0.38	
12	0.38	0.40	
13	0.40	0.43	
14	0.43	0.45	
15	0.45	0.47	
16	0.47	0.50	
17	0.50	0.52	
18	0.52	0.55	
19	0.55	0.58	
20	0.58	0.62	
21	0.62	$+\infty$	OF-Bin

Table 10.2: *Detector level binning in Q_e^2 (left) and y_e (right)*

$\mathcal{G}_{gen}$ Bin Grid Parameters	
Total number of bins n_{all}:	240
Overflow bins n_{OF}:	60
Low efficiency bins n_{LE} ($\epsilon < 70\%$):	74
Zero efficiency bins n_{ZE}:	0
Bins to unfold n:	240
Quotable bins n_{QO}:	106
Bins in Q^2 direction (without OF):	18
Bins in y direction (without OF):	10
$\mathcal{G}_{det}$ Bin Grid Parameters	
Total number of bins m_{all}:	836
Overflow bins m_{OF}:	116
Empty bins m_{EM}:	194
Bins to unfold m:	526
Bins in Q^2 direction (without OF):	36
Bins in y direction (without OF):	20

Table 10.3: *Properties of the used bin grids $\mathcal{G}_{gen}$ and $\mathcal{G}_{det}$.*
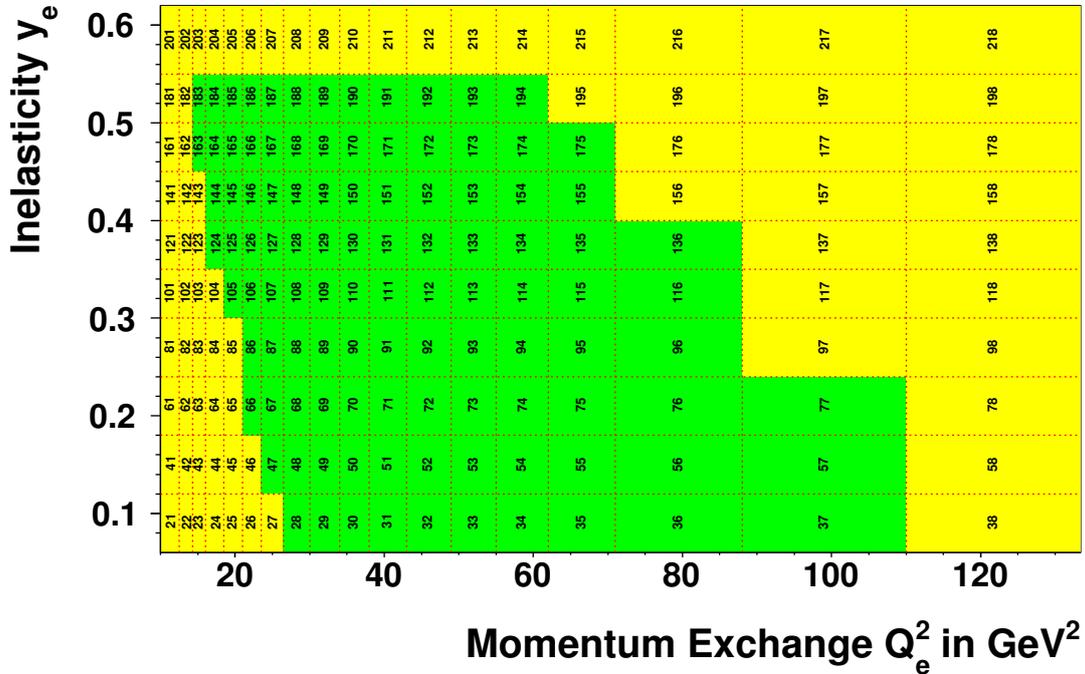
Figure 10.1.1: *Bin numbering scheme for the generator level bin grid $\mathcal{G}_{gen}$. Only the efficient bins which fulfill $\epsilon_{i_1,i_2} > \epsilon_{tr} = 70\%$ (green) will be quoted, all other bins (yellow) will be discarded after unfolding. Overflow bins are suppressed in this plot.*
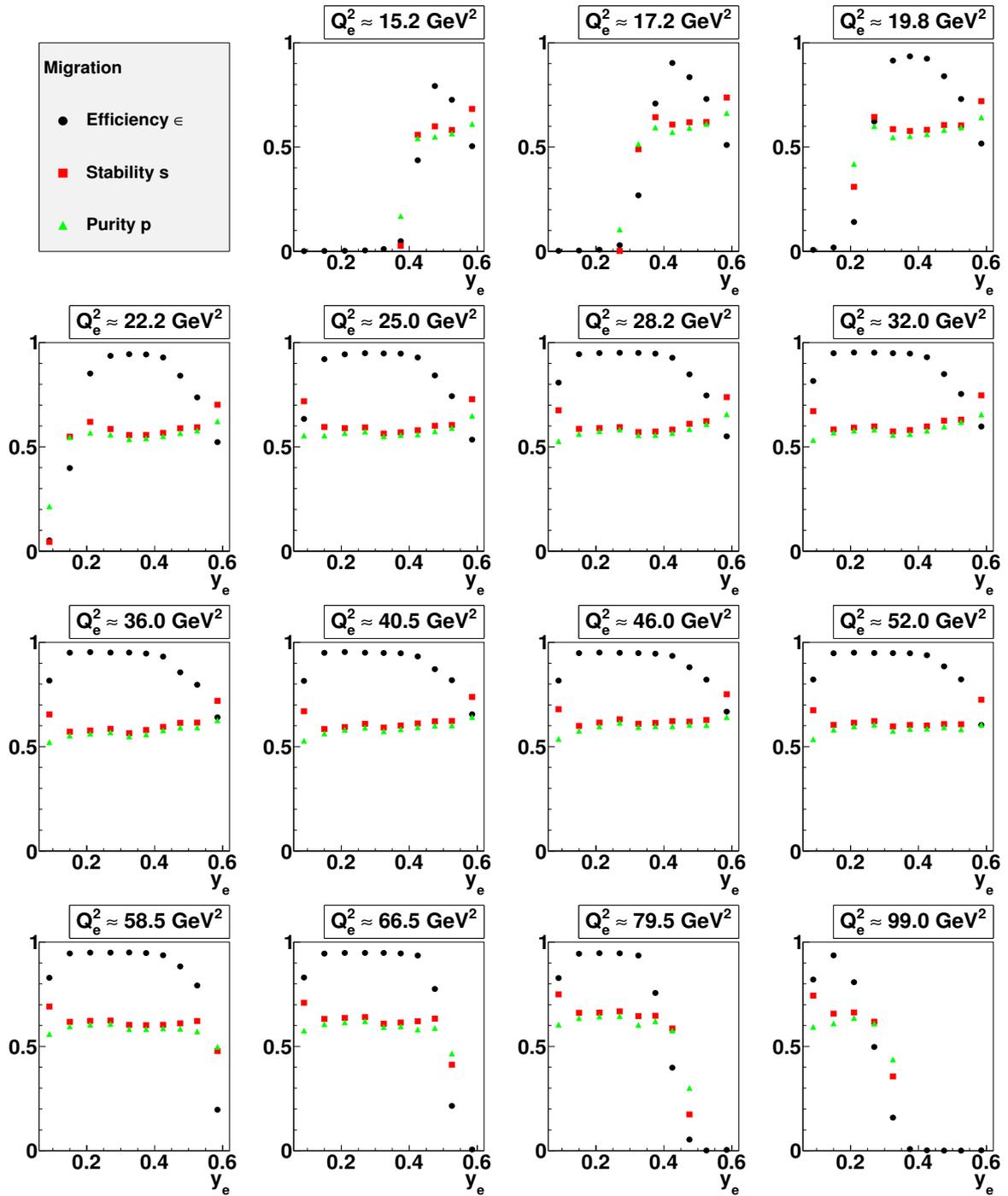
Figure 10.1.2: *Purity p_{i_1,i_2}, stability s_{i_1,i_2} and bin-by-bin effiency ϵ_{i_1,i_2}^{bbb}.*
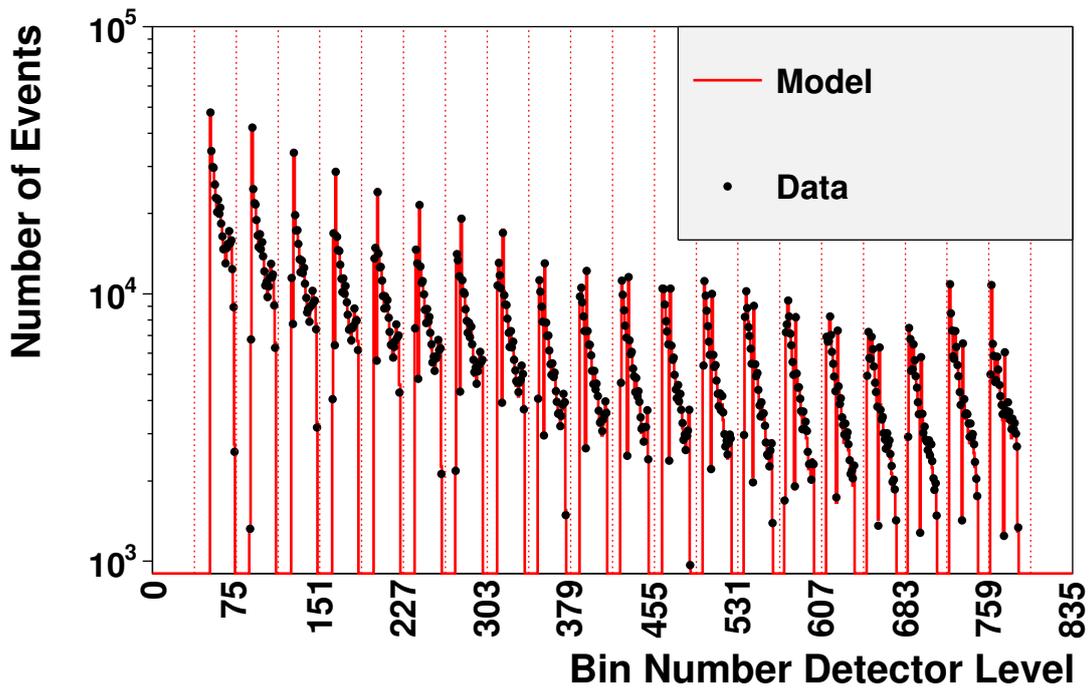
Figure 10.1.3: *Detector level distributions: Measured data $\vec{N}^{raw}$ events and recon-structed Monte Carlo events $\vec{N}^{rec}$ (*DJANGOH*). The model distributions is scaled to the data luminosity.*
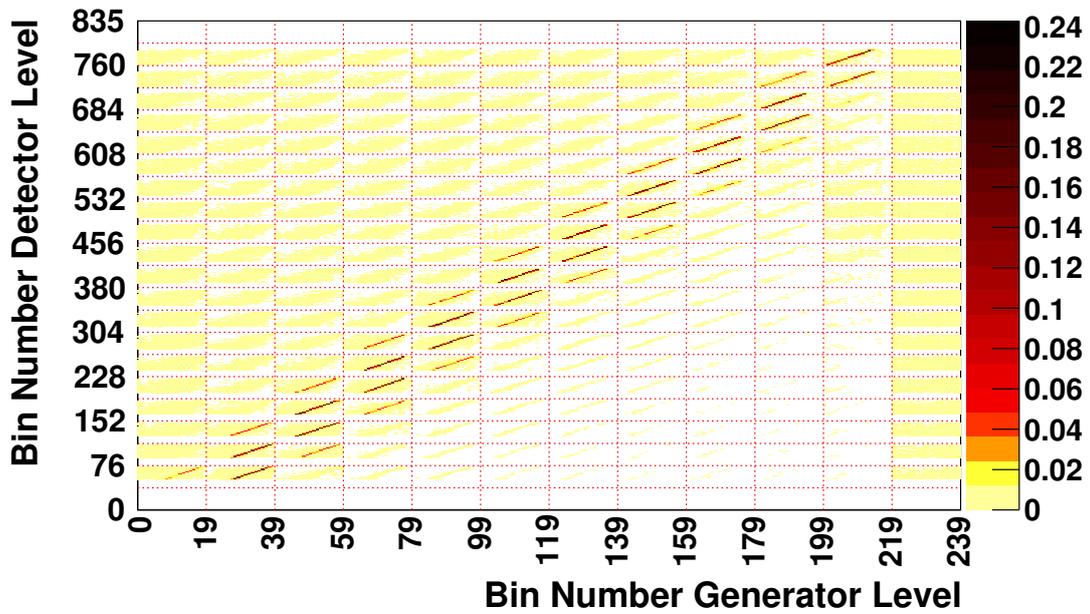
Figure 10.1.4: *Scaled response matrix $\mathbf{N}^{MC}$. For the bin numbering scheme, see figure (10.1.1).*
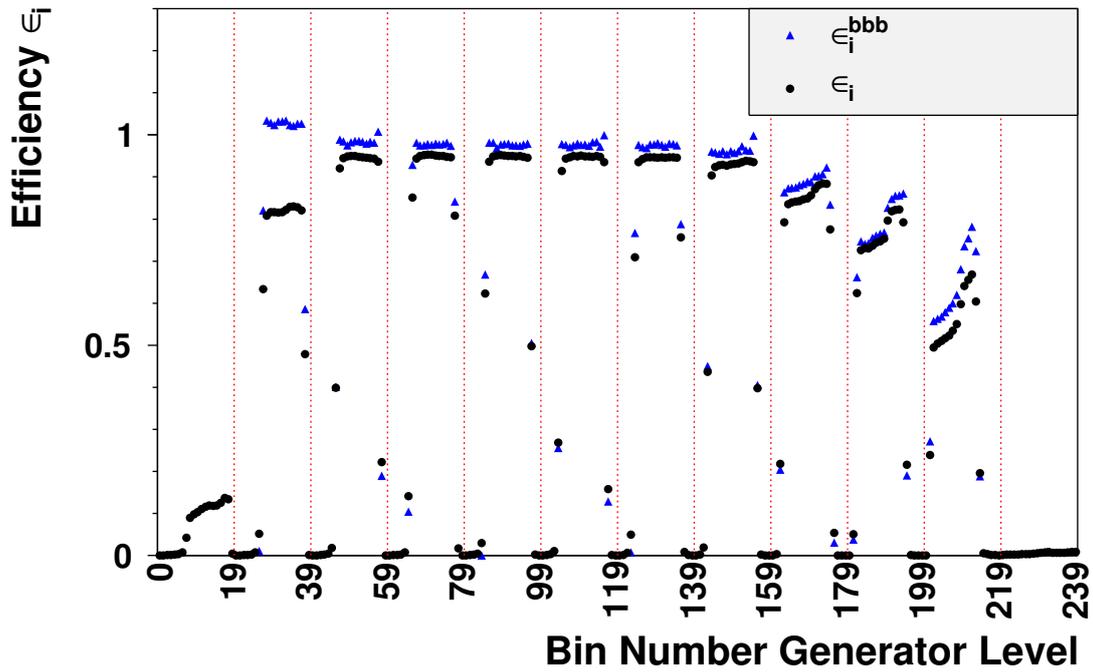
Figure 10.1.5: *Comparison of efficiencies. Shown are the overall efficiency ϵ_i, which is the product of the bin grid efficiency ϵ_i^{grid} and the reconstruction efficiency ϵ_i^{rec}, and the bin-by-bin efficiency ϵ_i^{bbb}. For the bin numbering scheme, see figure (10.1.1).*
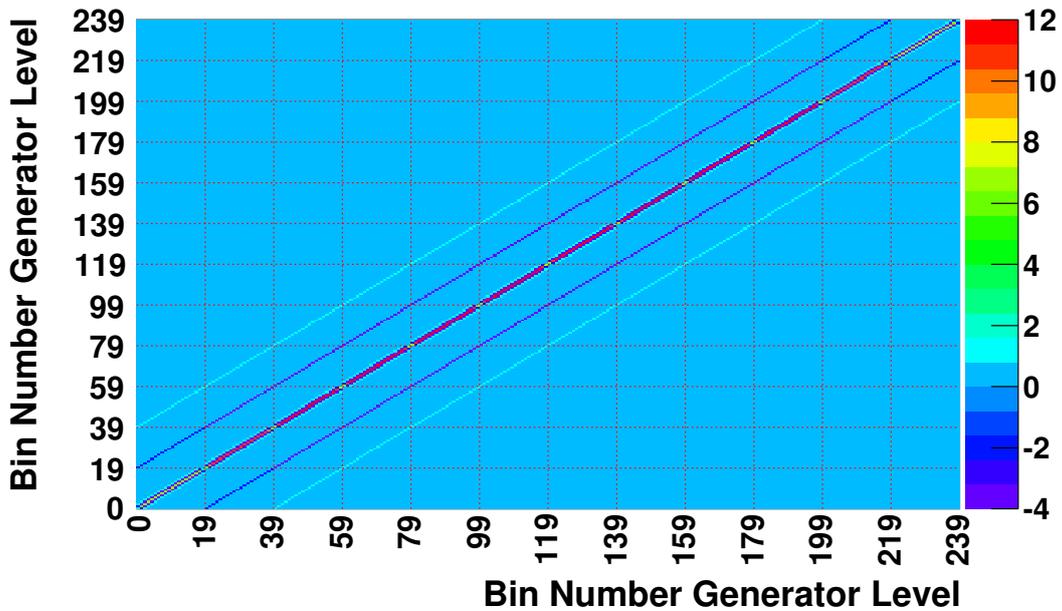
Figure 10.1.6: *The curvature matrix $\boldsymbol{L}^2$. For the bin numbering scheme, see figure (10.1.1).*
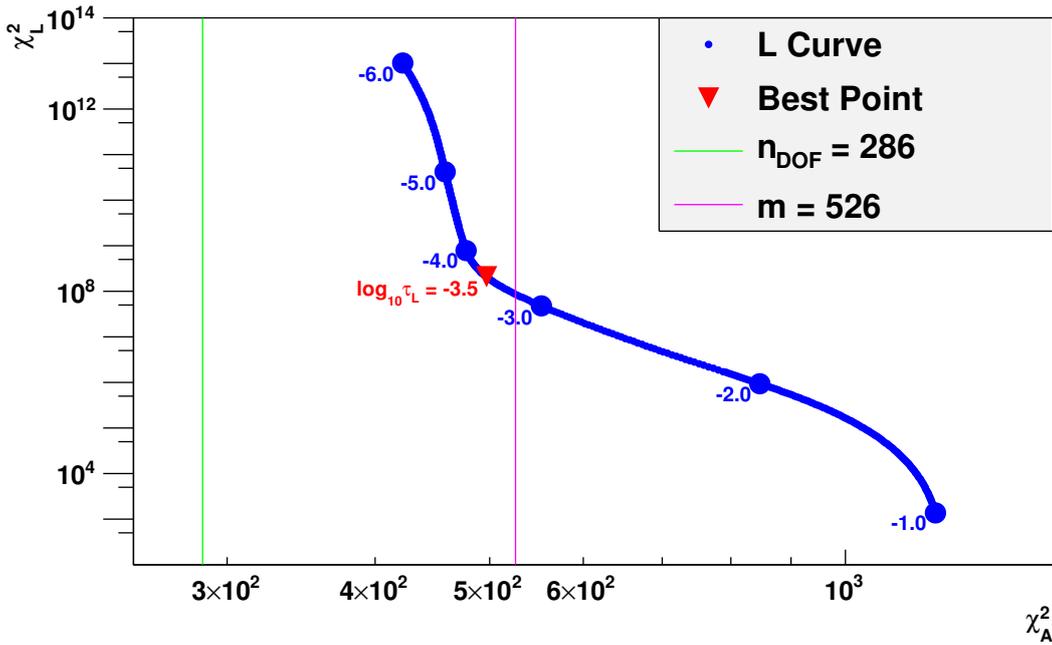
Figure 10.2.1: *L curve. The optimal point at* $\log \tau_L^2 = -3.46$ *is depicted. For orientation the positions of* n *and* n_{DOF} *are also indicated (see table 10.3).*

degrees of freedom. From that we obtain the ratio:

$$\frac{\chi_A^2}{n_{DOF}} \simeq 1.74$$

Note, however, that the n_{DOF} fit parameters are not statistically independent, such that this rather high value for χ_A^2/n_{DOF} seems reasonable.

10.2.2 Algebraic Analysis

We perform an algebraical analysis according to subsection 4.7.2. The matrix $\boldsymbol{E}_0$ (*Hesse matrix*) according to (4.2.7) is plotted in figure 10.2.2, while its eigenvalues are depicted in 10.2.6 (left). Note, how the eigenvalues of $\boldsymbol{E}_0$ span almost 12 orders of magnitude and exhibit a steep descend roughly after the 130th eigenvalue. This reflects the limited "rank" of the unfolding problem. Recall, that $\boldsymbol{E}_0$ corresponds to the inverse covariance matrix $\boldsymbol{COV}_{\vec{x}_0}^{-1}$ of the unregularized result $\vec{x}_0$.

Figure 10.2.6 also visualizes the curvature $\mathcal{C}_{curve}(\vec{a}_k)$ the modes $\vec{a}_k$, which are given according to the transformation $\boldsymbol{T} = \boldsymbol{T}_3 \boldsymbol{T}_2 \boldsymbol{T}_1$, see (4.4.13) and (4.4.12). Again, the curvature of the modes spans many orders of magnitudes. Note, that the first four modes are virtually flat. Since the curvature measure $\mathcal{C}_{curve}$ is only sensitive to the second derivative of any distribution $\vec{x}$, such modes are expected and correspond to the limited rank of the matrix $\boldsymbol{L}^2$. From mode 3 onwards, we observe physical modes slowly ascending in curvature. The higher the curvature, the slower the rise. From a mathematical point of view, this reflects the two-dimensionality of the problem, which introduces a degeneracy to the eigenspaces to $\boldsymbol{L}^2$. Roughly from mode 130 onwards, a
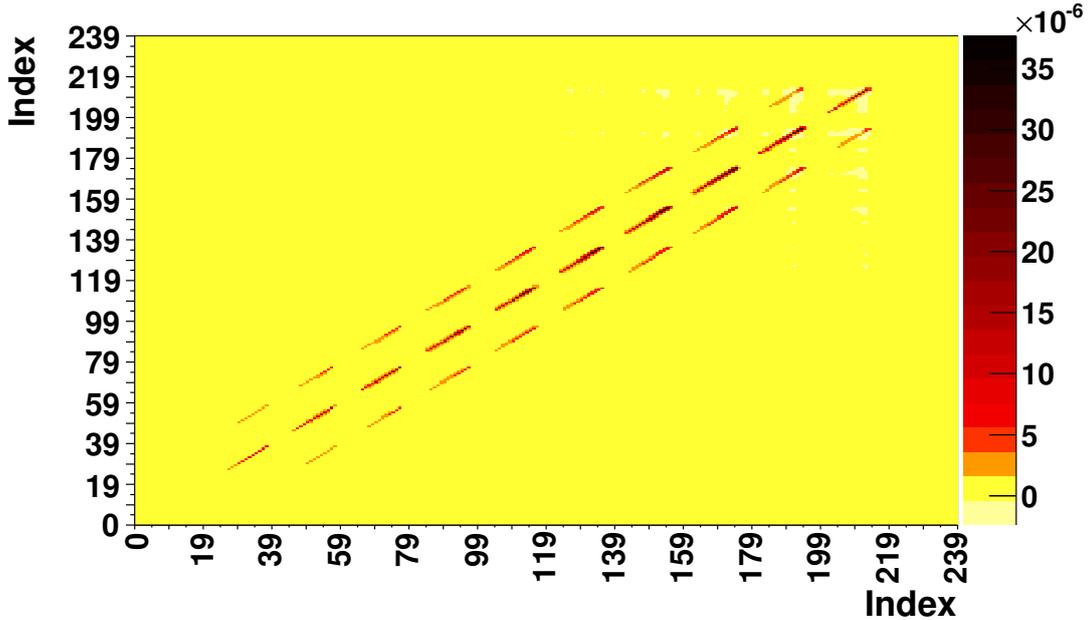
Figure 10.2.2: *The values of the matrix $\boldsymbol{E}_0$. It is clearly visible how it is dominated by its diagonal elements and those of neighbouring bins.*

sharp ascent is visible - modes beyond this kink are likely to be noise. Note, that the curvature matrix $\boldsymbol{L}^2$ transformes unter $\boldsymbol{T}$ to $\hat{\boldsymbol{L}}^2$, see figure 10.2.4.

The application of the transformation $\boldsymbol{T}$ is shown in figure 10.2.3. Both, the transformed unregularized unfolding result $\boldsymbol{T}\vec{x}_0$ and the transformed prediction $\boldsymbol{T}\vec{x}^{H1PDF2009}$ according to the pdf set `H1PDF2009` *[Kre, A$^+$09]* is drawn. Recall, that every coefficient $(\boldsymbol{T}\vec{x}_0)_i$ of the unregularized result has unit variance. The red line denotes the significance threshold $\sigma_{tr} = 3.84$ according to (4.4.15). The start of the noise region is clearly visible and is located after mode 145, leaving us with 146 significant modes (*effective dimension*):

$$n_0 = 146 \tag{10.2.3}$$

The boundary is depicted by the green line in figure 10.2.3. Note the region right from the green line: The smooth distribution according to the fit `H1PDF2009` continuously falls off, while the measured distribution shows noise fluctuations.

According to (4.4.19) and figure 10.2.5 we can determine the regularization parameter τ_A that corresponds to m_0. We obtain:

$$\log \tau_A \simeq -4.29 \tag{10.2.4}$$

10.2.3   Global Correlation Analysis

The global correlation analysis according to subsection 4.7.2 is done in figure 10.2.7. For the plot the averaged and the averaged squared reduced global correlation $\rho_{av}^{red}(\tau)$ and $\rho_{avsq}^{red}(\tau)$ have been drawn. As it has been discussed in subsection 4.4.3, only the bins with sufficient efficiency ($\epsilon > \epsilon_{tr} = 70\%$, quotable bins) have been used for the averaging process.
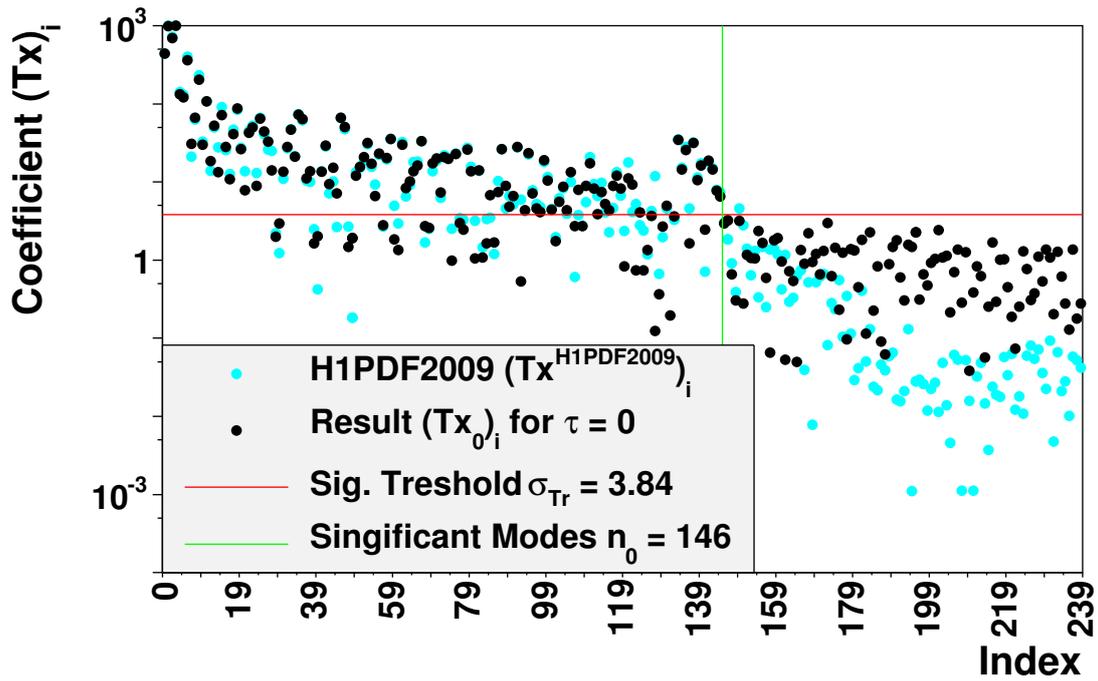
Figure 10.2.3: *Transformed unregularized result $\boldsymbol{T}\vec{x}_0$ and transformed prediction $\boldsymbol{T}\vec{x}^{H1PDF2009}$ according to the pdf set* H1PDF2009 *[Kre, A+09]. Transformation $\boldsymbol{T} = \boldsymbol{T}_3\boldsymbol{T}_2\boldsymbol{T}_1$ according to (4.4.12). Every coefficient $(\boldsymbol{T}\vec{x}_0)_i$ of the unregularized result has unit variance. The red line denotes the significance boundary $\sigma_{tr} = 3.84$ according to (4.4.15), the green line depicts the number of significant modes found (146).*
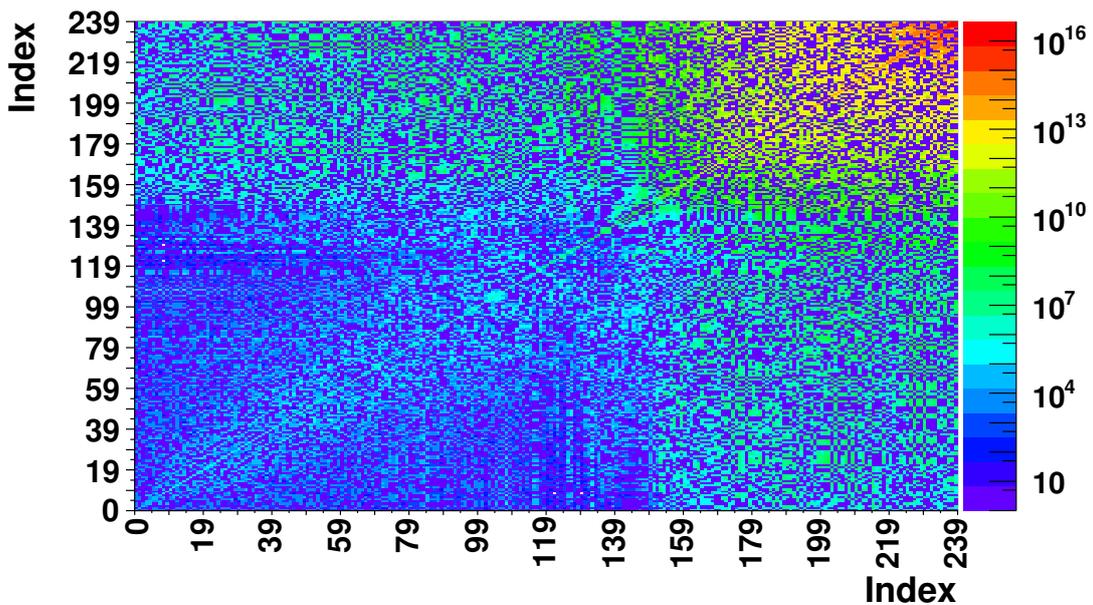
Figure 10.2.4: *Curvature matrix $\hat{\boldsymbol{L}}^2$, transformed according to (4.4.16).*
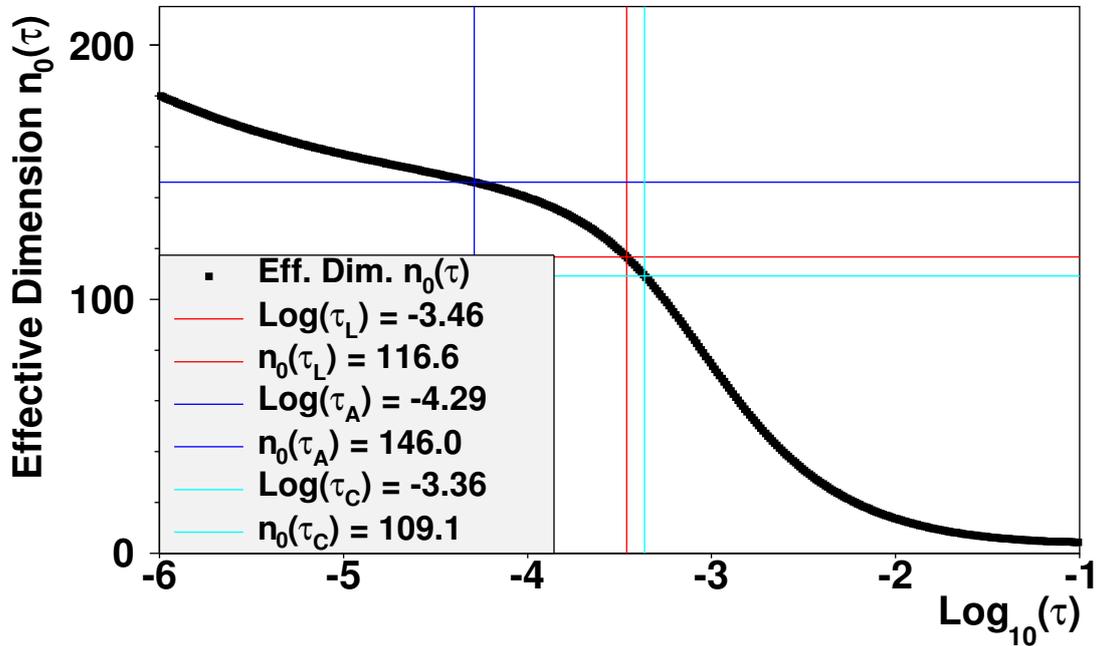
Figure 10.2.5: *Relation (4.4.19) between effective dimension n_0 and optimal regularization parameter τ_A. The values $\log \tau_A = -4.29$, $\log \tau_L = -3.46$ and $\log \tau_C = -3.36$ and its corresponding effective dimenions $n_0(\tau_A) \equiv n_0 = 146$, $n_0(\tau_L) = 116.6$, $n_0(\tau_C) = 109.2$ are depicted.*
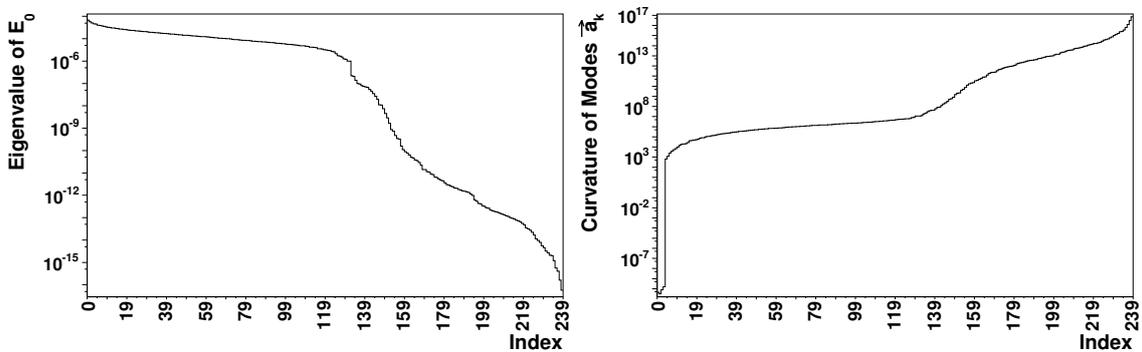
Figure 10.2.6: *Left: Eigenvalues of $\boldsymbol{E}_0$. Right: Curvature $\mathcal{C}(\vec{a}_k)$ of the modes $\vec{a}_k$.*

The minima in 10.2.7 are clearly visible. We obtain for the averaged squared reduced global correlation:

$$\log \tau_C \simeq -3.36 \qquad (10.2.5)$$

10.2.4 Interpretation and comparison

In order to further interpret the different propositions for τ, we can study how the unfolding result in a specific bin varies with τ. In figure 10.2.8 and 10.2.9 this is done for a single bin (bin 88), since from a single bin much can be learned in an exemplary manner. In particular, note that the τ-Axis can be divided in different regions, according to the functional form of the unfolding result $x_{88}(\tau)$, its statistical uncertainty $\sigma_{88}^{stat}(\tau)$ and the correlation to its neighbouring bins $\rho_{87,88}(\tau)$, $\rho_{89,88}(\tau)$, $\rho_{68,88}(\tau)$ and $\rho_{108,88}(\tau)$. Note, that bin 88 is located centrally in the efficient region of the phase space of this analysis (26.5 GeV^2 < Q^2 < 30 GeV^2 and 0.24 < y < 0.3).

For $\log \tau \lesssim -6$ the uncertainty $\sigma_{88}^{stat}(\tau)$ has a divergent behaviour for $\tau \to 0$, which mirrors the rank deficiency of the unfolding matrix $\boldsymbol{A}$. In this area, the result is dominated by unphysical fluctuations (*noise region*).

The region $-6 \lesssim \log \tau \lesssim -4$ exhibits a rather constant course of $\sigma_{88}^{stat}(\tau)$ (*plateau region*). The same is true for the absolute value $x_{88}(\tau)$ and the correlations to the neighbouring bins. In this region, rank deficiency problems have been remedied by a rather conservative choice of τ, while strong negative bin-to-bin correlations still remain. It comes at no surprise, that the value $\log \tau_A \simeq -4.29$ found from the algebraic analysis, belongs to the plateau region. Any choice of τ in this region is trustworthy in the sense that the result $x_{88}(\tau)$ and its error $\sigma_{88}^{stat}(\tau)$ will not depend too much on τ. Note, that the statistical uncertainty $\sigma_{88}^{stat}(\tau)$ is larger than the one from the bin-by-bin method σ_{88}^{bbb}. This is partly due to the negative correlations to the neighbouring bins.

The region around $-4 \lesssim \log \tau \lesssim -3.36 \simeq \log \tau_C$ shows a strong dependence of the error $\sigma_{88}^{stat}(\tau)$ on τ (*slope region*). Moreover, the correlations to the neighbouring bins dissapear while $\log \tau \to \log \tau_C$.

For $\log \tau \simeq \log \tau_C = -3.36$ we obtain a roughly diagonal covariance matrix $\boldsymbol{COV}_{\vec{x}}$ (*level of diagonalization*). Here, the reduced global correlation $\rho_{88}^{red}(\tau)$ has its minimum. Note, that at this point the statistical uncertainty $\sigma_{88}^{stat}(\tau_C)$ is free from contributions from negative correlations. Therefore, this value of $\sigma_{88}^{stat}(\tau_C)$ can be regarded as a measure for the "statistical accuracy" of this data point, see subsection 4.4.3. We see, that $\sigma_{88}^{stat}(\tau_C)$ is still larger than the uncertainty σ_{88}^{bbb} on the bin-by-bin result. We will come back to this observation in the next section.

Any $\log \tau \gtrsim \log \tau_C$ introduces strong positive correlations $\rho_{87,88}(\tau)$, $\rho_{89,88}(\tau)$, $\rho_{68,88}(\tau)$ and $\rho_{108,88}(\tau)$ to the neighbouring bins (*region of oversmoothing*). While the statistical uncertainties $\sigma_{88}^{stat}(\tau)$ are intriguingly small, the result $x_{88}(\tau)$ reflects rather the theoretical prediction x_{88}^{Theory} than the measurement. Any choice of τ from the oversmoothed region is therefore highly dangerous.

10.2.5 Choice of τ

In the context of this analysis the following viewpoint is adopted: Since the regularization should remove unphysical noise contributions while preserving the maximum of
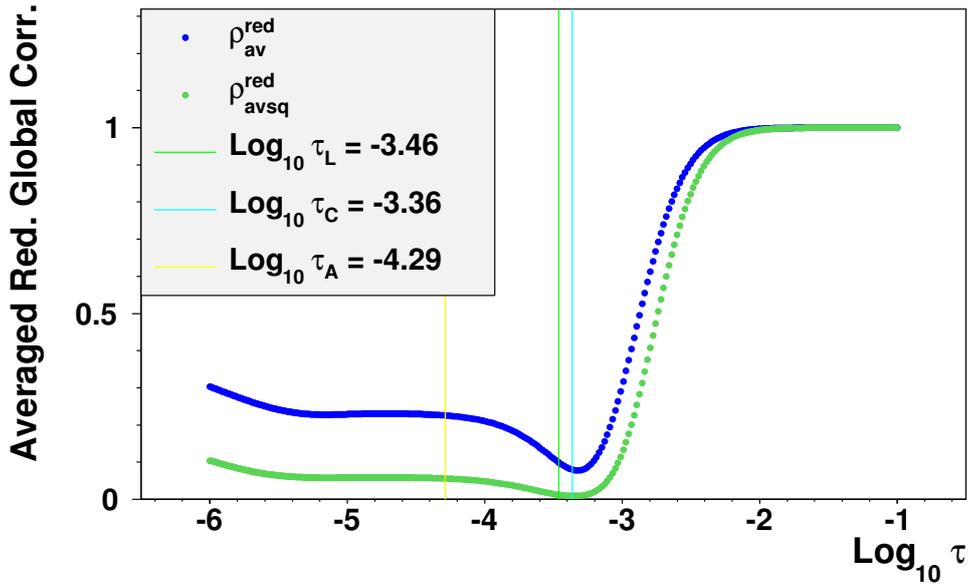
Figure 10.2.7: *Reduced global correlation $\rho_{av}(\tau)$ and reduced squared global correlation $\rho_{avsq}(\tau)$, averaged over all bins with sufficient efficiency ($\epsilon > \epsilon_{tr} = 70\%$, quotable bins). The vertical lines correspond to the values τ_A, τ_L and τ_C discussed in the text.*

physical information, a rather conservative choice of τ is made. Therefore, we choose

$$\log \tau_A = -4.29 \tag{10.2.6}$$

as the value for the regularization parameter.

10.3   The Unfolding Result

If not mentioned otherwise, the following results are all based on the choice (10.2.6) for τ.

10.3.1   The Result and its Statistical Uncertainties

In figures 10.3.1, 10.3.2 10.3.3 the effect of the unfolding procedure can be seen. Both the unfolding and the bin-by-bin method give consistent results. An important feature can be seen from figure 10.3.3, which shows that the relative error from unfolding is larger than the one from the bin-by-bin method.

There are two reasons for this behaviour. Firstly, for the conservative choice of $\log \tau_A \simeq -4.29$ the result exhibits negative correlations to neighbouring bins. These contribute to the statistical uncertainties, making them inconclusive at the same time.

The second reason seems to be more profound. In figure 10.3.4 the statistical errors of the result with the more progressive choice $\log \tau_C \simeq -3.36$ is compared to the errors of bin-by-bin unfolding. This level of regularization corresponds to the proposition of the global correlation method. As discussed in subsection 4.7.2, this regularization level gives statistical uncertainties, which can be viewed as a more conclusive measure for the statistical accuracy. We see, that the statistical uncertainties obtained from
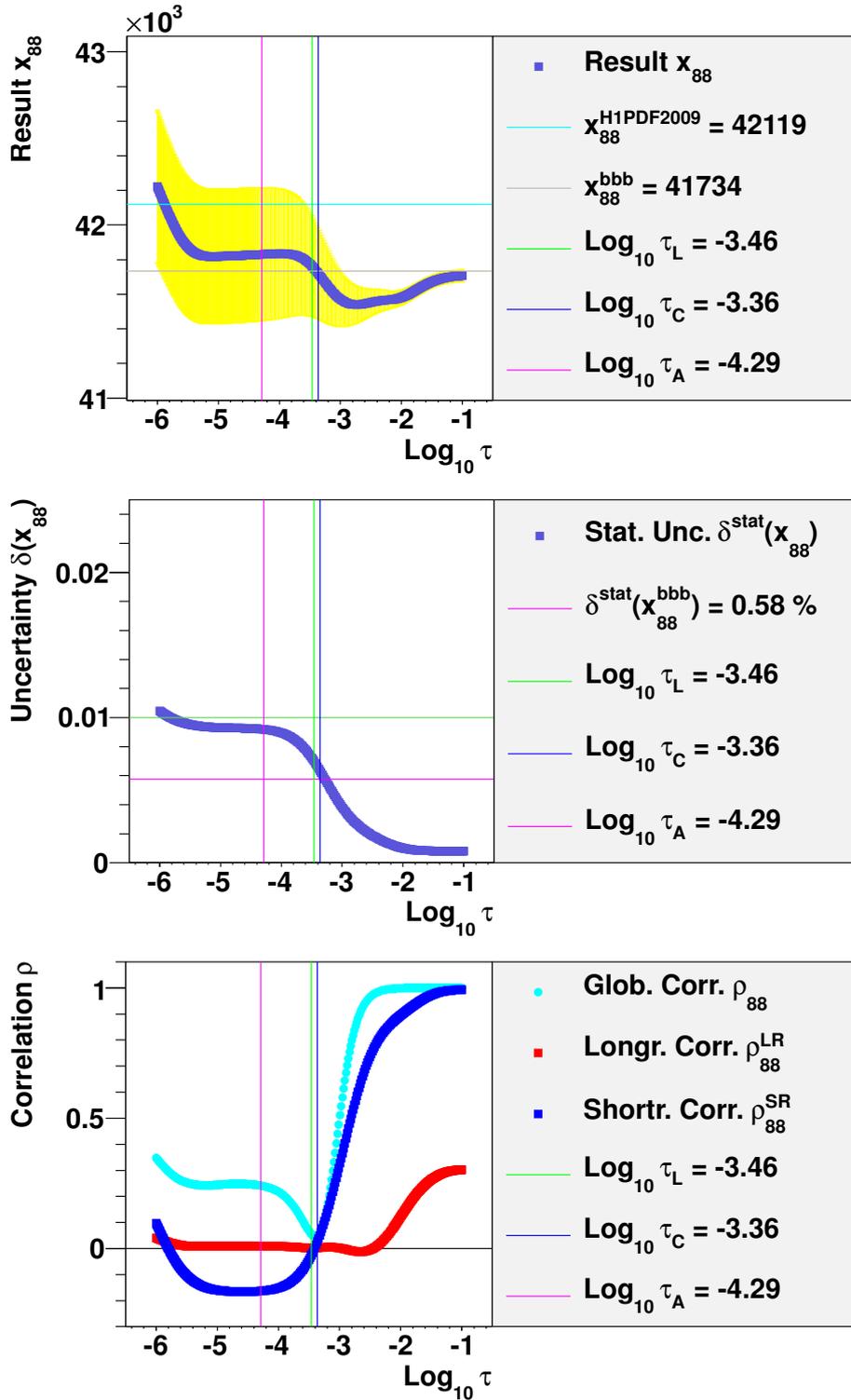
Figure 10.2.8: τ-dependence of unfolding results in bin 88 (26.5 $GeV^2 < Q^2 < 30$ GeV^2 and $0.24 < y < 0.3$). Top: Unfolding result $x_{88}(\tau)$ and result of bin-by-bin method (x_{88}^{bbb}). Center: Comparison of statistical uncertainties of both methods, $\delta^{stat}(x_{88}(\tau))$ and $\delta^{stat}(x_{88}^{bbb})$. Bottom: Reduced global correlation $\rho_{88}^{red}(\tau)$. Moreover, average correlations to the four nearest neighbours $(\rho_{88}^{SR}(\tau))$ and to all bins in the bin grid $(\rho_{88}^{SR}(\tau))$ are given.
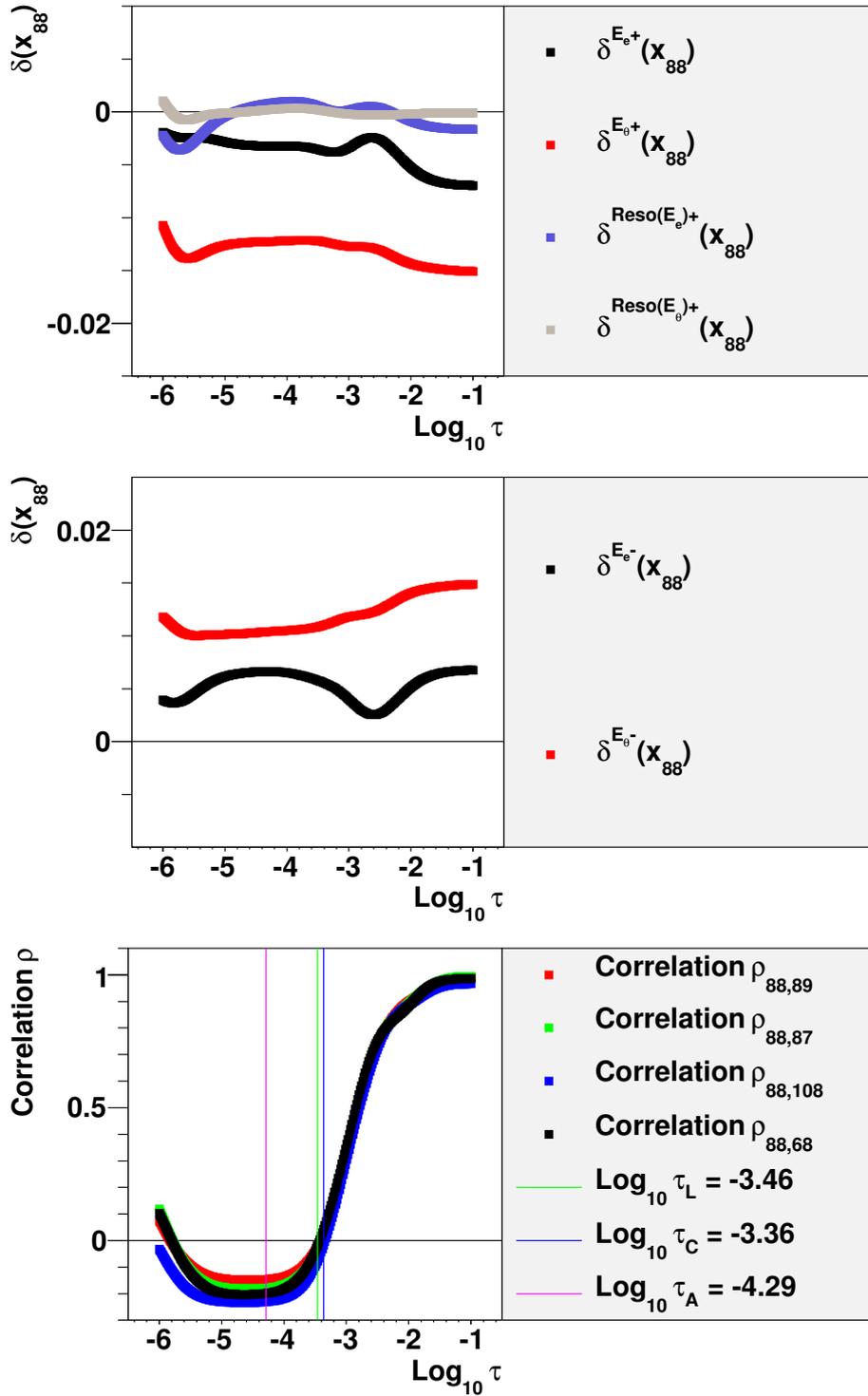
Figure 10.2.9: *τ-dependence of unfolding results in bin 88 (26.5 GeV² < Q² < 30 GeV² and 0.24 < y < 0.3). Top and center: Resulting systematic shifts $\delta^\lambda (x_{88})$ on the result for different systematic error sources λ. We only show the most prominent error sources resulting from the resolution and calibration of the electron angle θ_e and the electron energy E_e. All other shifts are on a permille scale. Bottom: Correlation to the four nearest neighbours $\rho_{87,88}(\tau)$, $\rho_{89,88}(\tau)$, $\rho_{68,88}(\tau)$ and $\rho_{108,88}(\tau)$.*

unfolding are still higher than those from the bin-by-bin method! The discrepancy takes on values up to 30 %. Obviously, the bin-by-bin method underestimates the statistical uncertainties in this case.

In figure 10.3.6 and 10.3.7 the full correlation matrix $\boldsymbol{Corr_{\vec{x}}}$ and the reduced global correlation ρ_i is plotted. Note the non-vanishing off-diagonal elements of $\boldsymbol{Corr_{\vec{x}}}$.
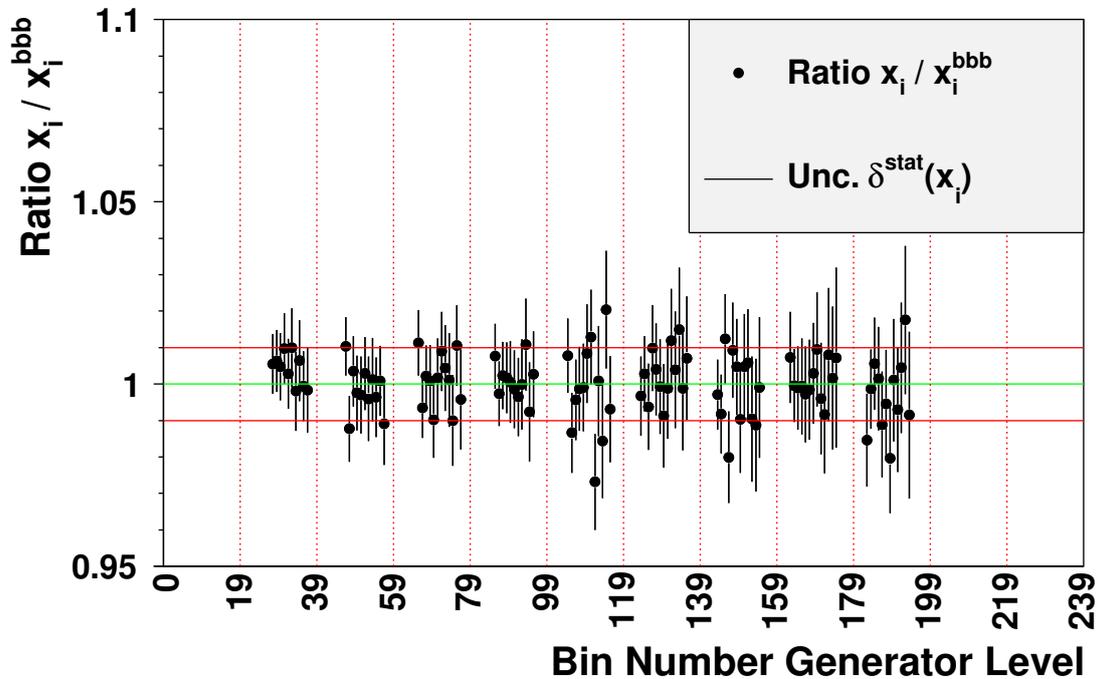
Figure 10.3.1: *Bin wise ratio $\frac{x_i}{x_i^{bbb}}$ of the unfolding result $\vec{x}$ and the result of the bin-by-bin method $\vec{x}^{bbb}$. The error bars correspond to the statistical error $\delta^{stat}(x_i)$ of the unfolding result. Level of regularization: $\log\tau_A \simeq -4.29$. For the bin numbering scheme, see figure (10.1.1).*
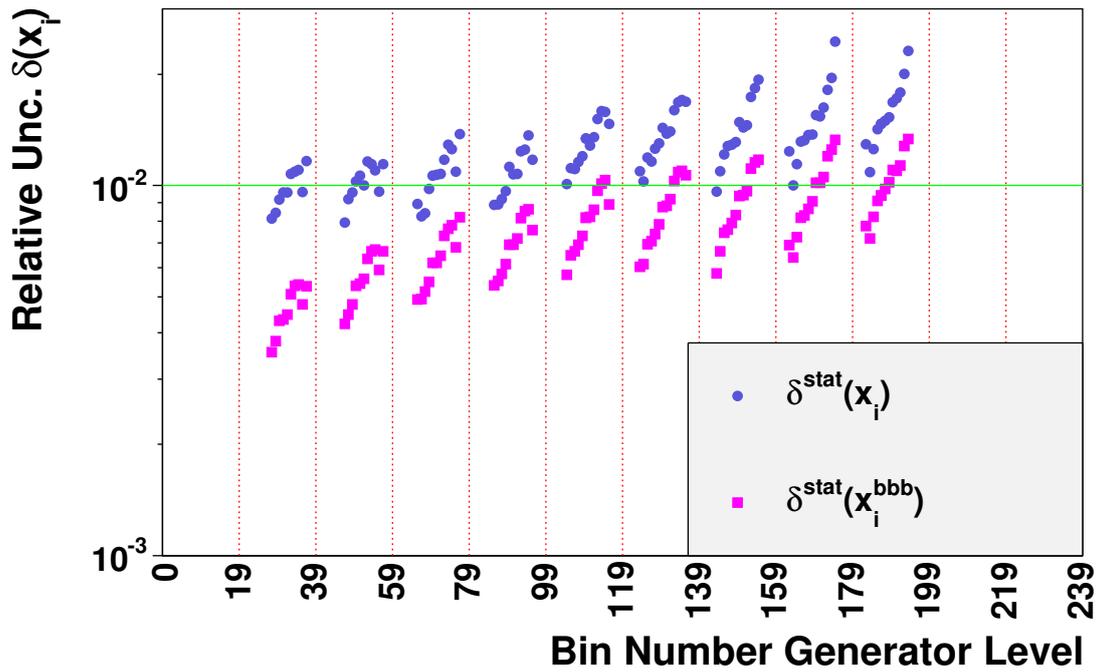
Figure 10.3.2: *Statistical uncertainties $\delta^{stat}(x_i)$ and $\delta^{stat}\left(x_i^{bbb}\right)$ of the unfolding result $\vec{x}$ and the result from the bin-by-bin method $\vec{x}^{bbb}$ are drawn. Level of regularization: $\log \tau_A \simeq -4.29$. For the bin numbering scheme, see figure (10.1.1).*
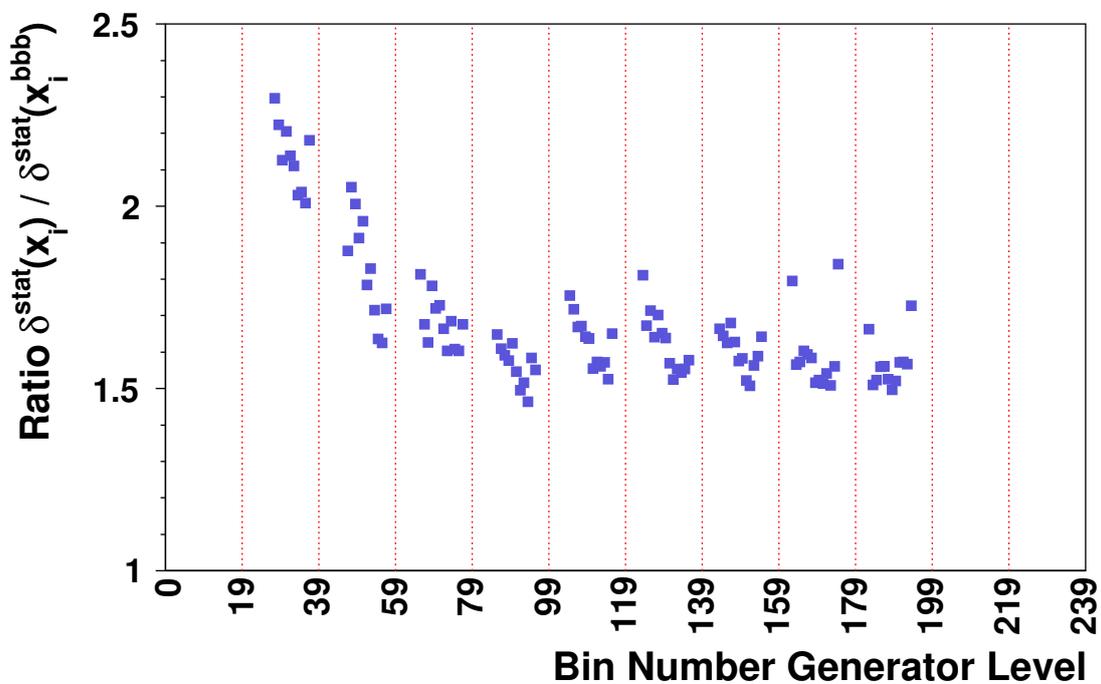
Figure 10.3.3: *Ratio $\frac{\delta^{stat}(x_i)}{\delta^{stat}\left(x_i^{bbb}\right)}$ of statistical uncertainties of unfolding method $\delta^{stat}(x_i)$ and bin-by-bin method $\delta^{stat}\left(x_i^{bbb}\right)$. Level of regularization: $\log \tau_A \simeq -4.29$. For the bin numbering scheme, see figure (10.1.1).*
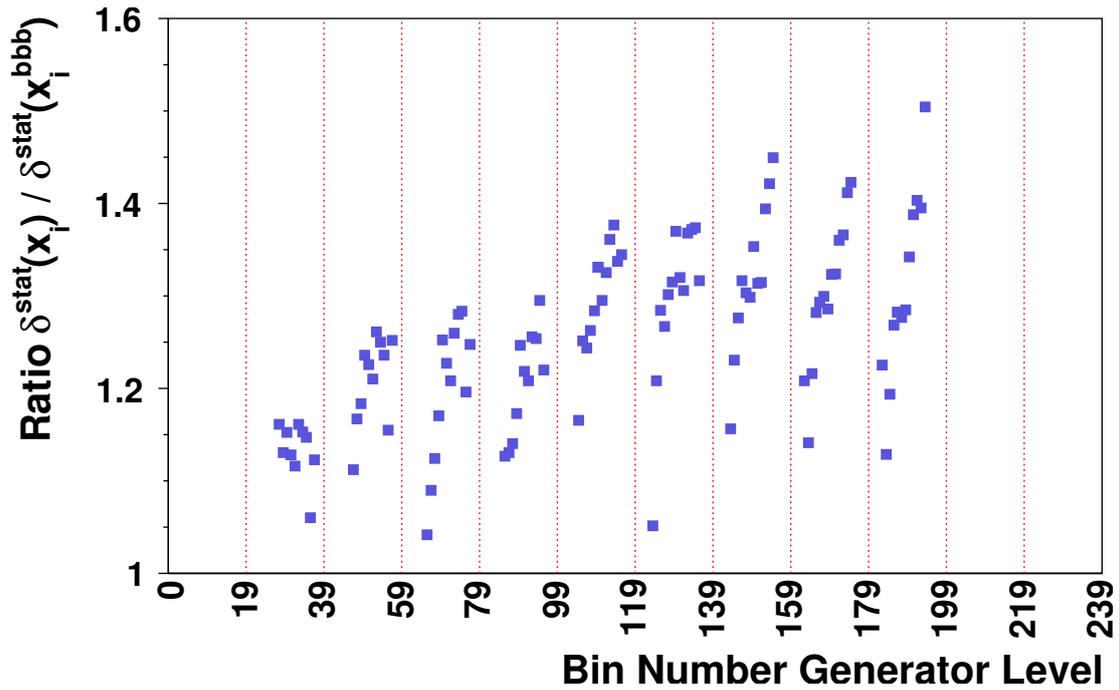
Figure 10.3.4: *Ratio* $\frac{\delta^{stat}(x_i)}{\delta^{stat}\left(x_i^{bbb}\right)}$ *of statistical uncertainties of unfolding method* $\delta^{stat}(x_i)$ *and bin-by-bin method* $\delta^{stat}\left(x_i^{bbb}\right)$. *Level of regularization:* $\log\tau \simeq -3.4$. *For the bin numbering scheme, see figure (10.1.1).*
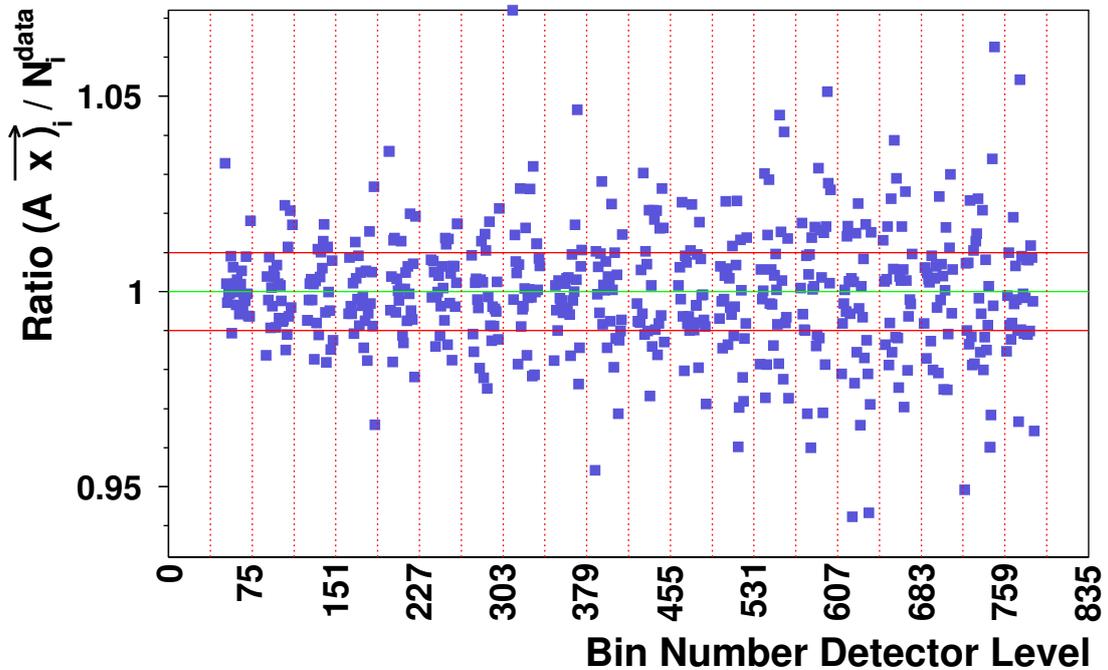
Figure 10.3.5: *Bin wise ratio* $\frac{(\boldsymbol{A}\vec{x})_i}{N_i^{data}}$ *of the result folded back* $\boldsymbol{A}\vec{x}$ *and the measured data* $\vec{N}^{data}$ *distribution. Level of regularization:* $\log\tau_A \simeq -4.29$.
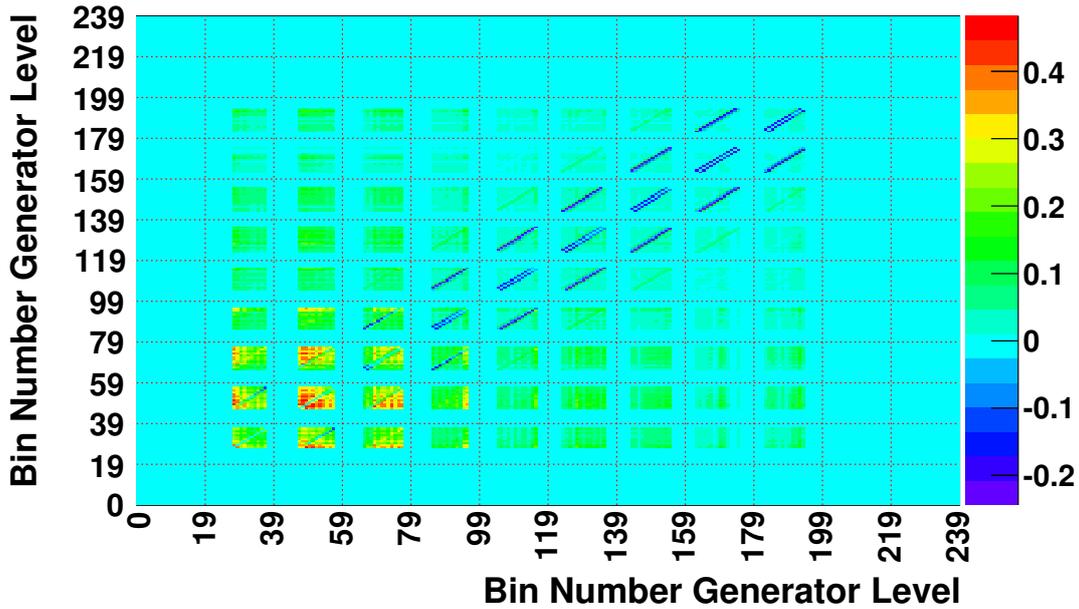
Figure 10.3.6: *Full matrix* $\boldsymbol{Corr}_{\vec{x}}^{stat}$ *of statistical correlations. Level of regularization:* $\log \tau_A \simeq -4.29$. *For the bin numbering scheme, see figure (10.1.1).*
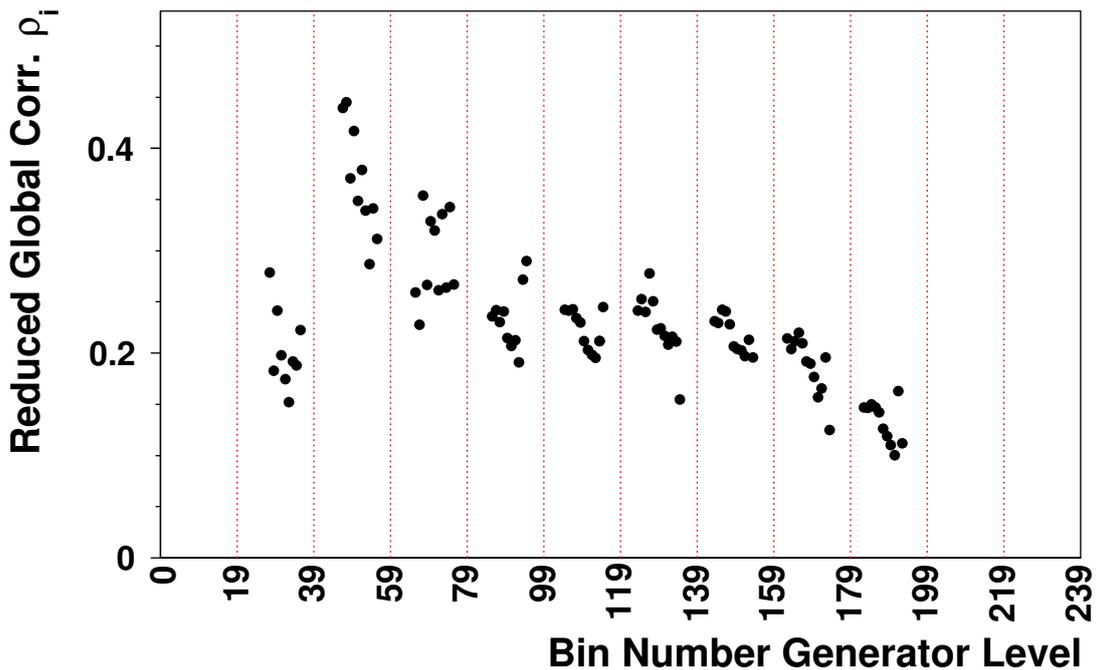
Figure 10.3.7: *Reduced global correlation* ρ_i^{red} *for each bin. Level of regularization:* $\log \tau_A \simeq -4.29$. *For the bin numbering scheme, see figure (10.1.1).*

10.3.2   Background Systematics

In figure 10.3.8 the different contributions to the statistical uncertainty $\delta^{stat}(x_i)$ according to (4.8.1) and the discussion in subsection 4.8.1 are drawn. The plot shows

the statistical uncertainty from the raw measurement $\delta^{raw}(x_i)$, from the background statistics $\delta^{bgr}(x_i)$ and from the background scale uncertainty $\delta^{bsc}(x_i)$.

Note, that for the background scale uncertainty the rather conservative choice of $\delta(f_{bgr}) = 0.5$ was made, see (7.2.4). This reflects the difficulty to determine the background normalization scale discussed in chapter 7. However, due to the low background contamination neither the statistical uncertainty $\delta^{bgr}(x_i)$ nor the scale uncertainty $\delta^{bsc}(x_i)$ of the background play a dominant role in the overall uncertainty.
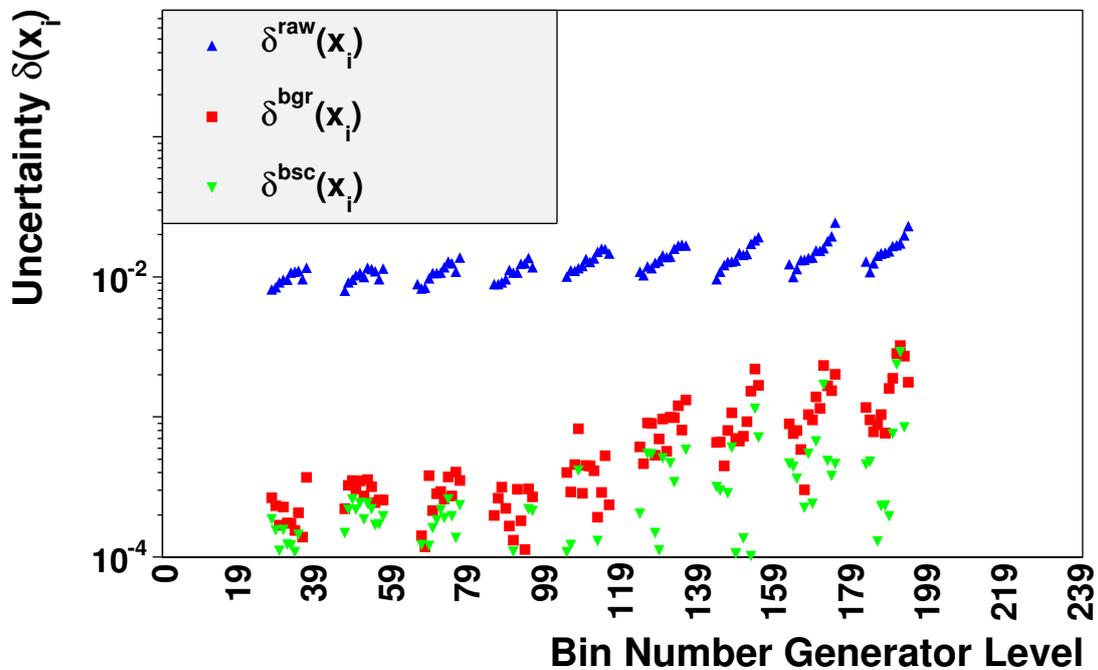
Figure 10.3.8: *Statistical uncertainty $\delta^{raw}(x_i)$ of the raw measurement, $\delta^{bgr}(x_i)$ from the background statistics and $\delta^{bsc}(x_i)$ from the background scale uncertainty. For the bin numbering scheme, see figure (10.1.1).*
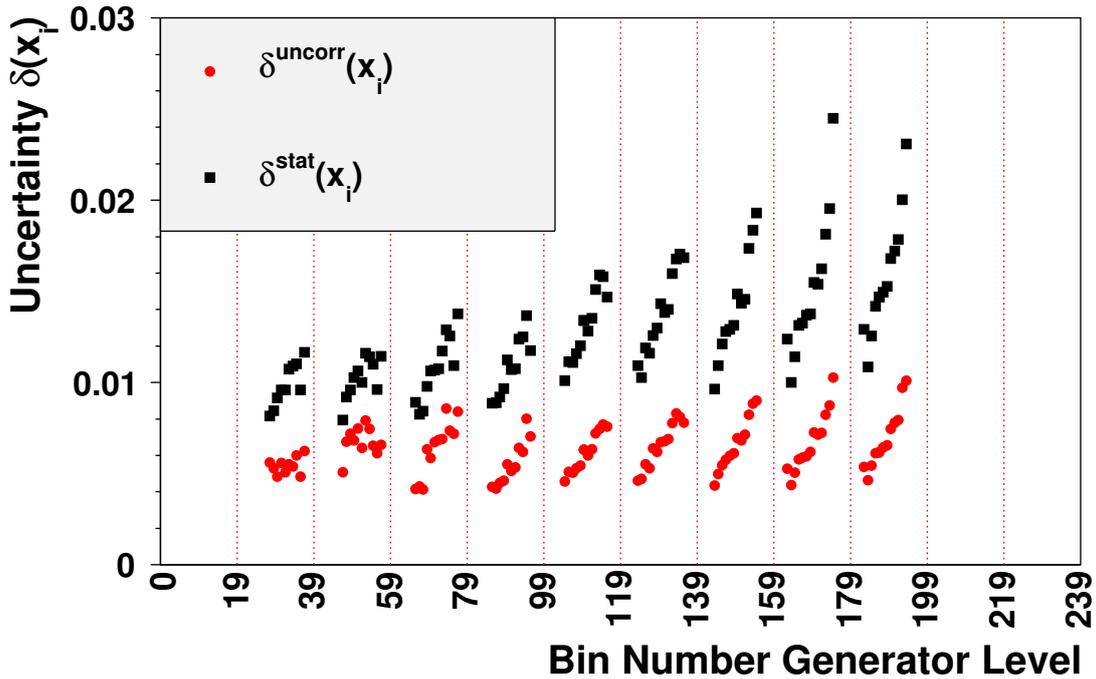
Figure 10.3.9: *Uncorrelated systematic uncertainty $\delta^{uncorr}(x_i)$ according to (4.8.12) and statistical uncertainties $\delta^{stat}(x_i)$.   For the bin numbering scheme, see figure (10.1.1).*

10.3.3   Uncorrelated Systematics

As discussed in subsection 4.8.1, the statistical uncertainties of the Monte Carlo model introduce a systematic uncertainty, which is determined according to (4.8.12). The uncorrelated systematic uncertainties $\delta^{uncorr}(x_i)$ are shown in in figure 10.3.9, together with the statistical uncertainty $\delta^{stat}(x_i)$ for comparison. Due to the high luminosity of the Monte Carlo simulation, the contribution of the uncorrelated systematic uncertainty to the overall error is rather small, taking on values on the permille scale throughout the phase space.

10.3.4   Correlated Systematics

In this subsection the different sources of correlated systematic uncertainties are propagated to the result. This is done according to subsection 4.8.3, using three independent response matrices $N_{ji}^{+\lambda}$, N_{ji}^{MC} and $N_{ji}^{-\lambda}$ for each error source λ. This provides us with two relative shifts $\delta^{\lambda+}(x_i)$ and $\delta^{\lambda+}(x_i)$ on the result x_i in each bin. Exceptional are the energy resolution uncertainty $\delta^{Res(E_e)}(x_i)$ and the polar angle resolution uncertainty $\delta^{Res(\theta_e)}(x_i)$, for which only two independent response matrices $N_{ji}^{Res(E_e)}$, N_{ji}^{MC} and $N_{ij}^{Res(\theta_e)}$, N_{ji}^{MC} are compared, providing only one shift for each error source.

   In figure 10.3.10 to 10.3.25, the procedure is done for the various error sources. For each source, the systematic shifts $\delta^{\lambda\pm}(x_i)$ according to the propagation method (4.8.19) is calculated, together with the result $\delta_{offset}^{\lambda\pm}(x_i)$ of the offset method (4.8.20) and the result $\delta_{bbb}^{\lambda\pm}(x_i)$ of the bin-by-bin method (4.8.21) for comparison. While the propagation method and the offset method give comparable results for every error

source λ, they both seem to differ from the bin-by-bin method in some cases. This is remarkable, since it casts doubt on the systematic error treatment of the bin-by-bin method in the presence of strong migration effects.

Note, that for the final result, only the result $\delta^{\lambda\pm}(x_i)$ of the propagation method will be used. Moreover, with the exeption of the resolution uncertainties $\delta^{Res(E_e)}(x_i)$ and $\delta^{Res(\theta_e)}(x_i)$, the systematic shifts will be symmetrized for each source λ, according to the symmetrization prescription (4.8.22).

Electron Energy Scale. According to chapter 9, the matrices $N_{ji}^{E_e\pm}$ have been produced using an input shift of $(\Delta E_e)_\pm = \pm 50$ MeV. From figure 10.3.10 and 10.3.11 it is obvious, that the systematic uncertainty $\delta^{E_e\pm}(x_i)$ dominates the region of small y_e, corresponding to high values of E_e (close to the kinematic peak $E_e \simeq 27.6$ GeV). In this area, $\delta^{E_e\pm}(x_i)$ takes on values clearly above 1 %. This is a consequence of the migration effects, which occur due to the shift $(\Delta E_e)_\pm$ in the energy scale and which predominantly affect the regions of steep y_e-dependance. Note, that for small y, the systematic uncertainty from the propagation method is significantly smaller than that from the bin-by-bin method.

Electron Polar Angle Scale. According to chapter 9, the matrices $N_{ji}^{\theta_e\pm}$ have been produced using an input shift of $(\Delta\theta_e)_\pm = \pm 1$ mrad. From 10.3.12 and 10.3.13 it can be seen, that the systematic uncertainty $\delta^{\theta_e\pm}(x_i)$ dominates the region of small Q_e^2, corresponding to high values of θ_e (close to the beam pipe at $\theta_{pipe} \simeq \pi$). In this area, $\delta^{\theta_e\pm}(x_i)$ takes on values clearly around 1 to 2 %. From all systematic error sources, this seems to be the dominating one in most regions. The shape of $\delta^{\theta_e\pm}(x_i)$ is again a consequence of migration effects, which occur due to the shift $(\Delta\theta_e)_\pm$ in the polar angle scale and which predominantly affect the regions of steep Q_e^2-dependance. Note, that for small y_e, the systematic uncertainty from the propagation method is typically larger than that from the bin-by-bin method.

Hadron Energy Scale. The effect of the energy scale of the hadronic final state (HFS) is expected to be rather small, since it affects the reconstructed events only through the $(E - p_z)_{tot}$-cut (6.3.8). However, the error propagation has been performed using a shift of $(\Delta E_{HFS})_\pm = \pm 4\% \cdot E_{HFS}$ as input for the simulation of $N_{ji}^{E_{HFS}\pm}$. Note in 10.3.14 and 10.3.15, how the shift $(\Delta E_{HFS})_\pm$ indeed slightly affects the data selection, favoring the low-y_e-region for positive shifts ΔE_{HFS}. The effect is on a scale of 2 permille maximum.

Cip Efficiency. The matrices $N_{ji}^{w_{Cip}\pm}$ have been produced applying the shifted weight functions $w^{Cip\pm}$ from chapter 8, see figure 8.2.4. Since the Cip Hit Cut is by construction constrained to the region $E_e < 16$ GeV (see chapter 6), a significant uncertainty occurs only in the three highest Q_e^2-bins, see 10.3.16 and 10.3.17. The effect is on a scale of 4 permille maximum. The small systematic uncertainty in the low-Q_e^2-region is an artefact of the regularization procedure.

Ecra Cut Efficiency. The matrices $N_{ji}^{w_{ECRA}\pm}$ have been produced applying the shifted weight functions $w^{ECRA\pm}$ from subsection 8, see figure 8.2.9. Recall, that the ECRA cut is a function of E_e and R_e being particularly restrictive at small energies and small radii. For $R_e > 55$ cm and $E_e > 18$ GeV it is relaxed completely. This is reflected in figure 10.3.18 and 10.3.19, which shows, that significant uncertainties occur only for high y_e and low Q_e^2. In this region, however, the effect is on a scale of roughly 1.2 %. The small systematic uncertainties in the low-y_e-region is an artefact of the regularization procedure.

Vertex Reconstruction Efficiency. The matrices $N_{ji}^{w_{VtxEff}\pm}$ have been produced applying the shifted weight functions $w^{VtxEff\pm}$ from subsection 8, see figure 8.2.7. Recall, that the combined vertex efficiency ϵ^{VtxEff} from the CJC vertex reconstruction and the CIP vertex reconstruction is very high, exhibiting a remaining inefficiency on the sub permille scale. This is reflected in figures 10.3.20 and 10.3.21, too.

S3 Trigger Efficiency. The matrices $N_{ji}^{w_{Trg}\pm}$ have been produced applying the shifted weight functions $w^{Trg\pm}$ from subsection 8, see figure 8.2.2. The result is shown in 10.3.22 and 10.3.23. The effect varies mit Q_e^2, reflecting the slight deficiencies in the S3 trigger description at very low and very high SpaCal radii. The uncertainty adopts values between 1 and 3 permille. Notice the asymmetry in the shifts $\delta^{Trg+}(x_i)$ and $\delta^{Trg-}(x_i)$. This is an artefakt of the high overall S3 trigger effiency, which imposes a boundary condition on the upshifted weight function $w^{Trg\pm}$, see chapter 8.

Electron Resolution Uncertainty. The matrix $N_{ji}^{Res(E_e)}$ has been produced by a variation of the electron energy resolution. This is done by implementing an additional smearing of the reconstructed electron energy with a GAUSS-distribution of variance $\delta_{E_e}^{GAUSS} = 200$ MeV. Recall from chapter 9, that we have $Res(E_e) \simeq 1$ GeV troughout the phase space. Therefore, the GAUSS-smearing effectively widens the electron resolution by:

$$\Delta Res(E_e) = \sqrt{(Res(E_e))^2 + \left(\delta_{E_e}^{GAUSS}\right)^2} - Res(E_e) \simeq 40 \text{ MeV} \qquad (10.3.1)$$

This number reflects the value motivated in (9.1.3) and (9.1.5).

As an effect of this resolution variation we obtain randomly distributed statistical fluctuations, see figure 10.3.24. This is not a surprise, since an increase in smearing is expected to enhance the bin-by-bin correlations and therefore the statistical uncertainties. Note, however, that for small y_e, there seems to be a small negative systematic shift. This can be viewed as an "over-correction" of migration effects done by an unfolding procedure with the shifted matrix $N_{ji}^{Res(E_e)}$. With few exceptions, all shifts and fluctuations are on a scale signifiantly below 1 %.

Polar Angle Resolution Uncertainty. The matrix $N_{ji}^{Res(\theta_e)}$ has been produced by a variation of the electron polar angle resolution. This is done by implementing an additional smearing of the reconstructed polar angle with a GAUSS-distribution of variance $\delta_{\theta_e}^{GAUSS} = 1.83$ mrad. Recall from chapter 9, that we have roughly $Res(\theta_e) \simeq$ 4 mrad troughout the phase space. Therefore, the GAUSS-smearing effectively widens the electron resolution by:

$$\Delta Res(\theta_e) = \sqrt{(Res(\theta_e))^2 + \left(\delta_{\theta_e}^{GAUSS}\right)^2} - Res(\theta_e) \simeq 0.4 \text{ mrad} \qquad (10.3.2)$$

This number reflects the value motivated in (9.2.3) and (9.2.5).

As an effect of this scale variation we again obtain randomly distributed statistical fluctuations, see figure 10.3.25. All shifts and fluctuations are on a scale signifiantly below 1 %.

Vertex Shape Reweighting Uncertainty. As pointed out in chapter 8, a z Vertex shape reweighting has been applied to the Monte Carlo simulation. The matrices $N_{ji}^{w_{VtxRew}\pm}$ have been produced applying the shifted weight functions $w^{VtxRew\pm}$ from chapter 8, see figure 8.3.1. The result is shown in 10.3.27 and 10.3.27.
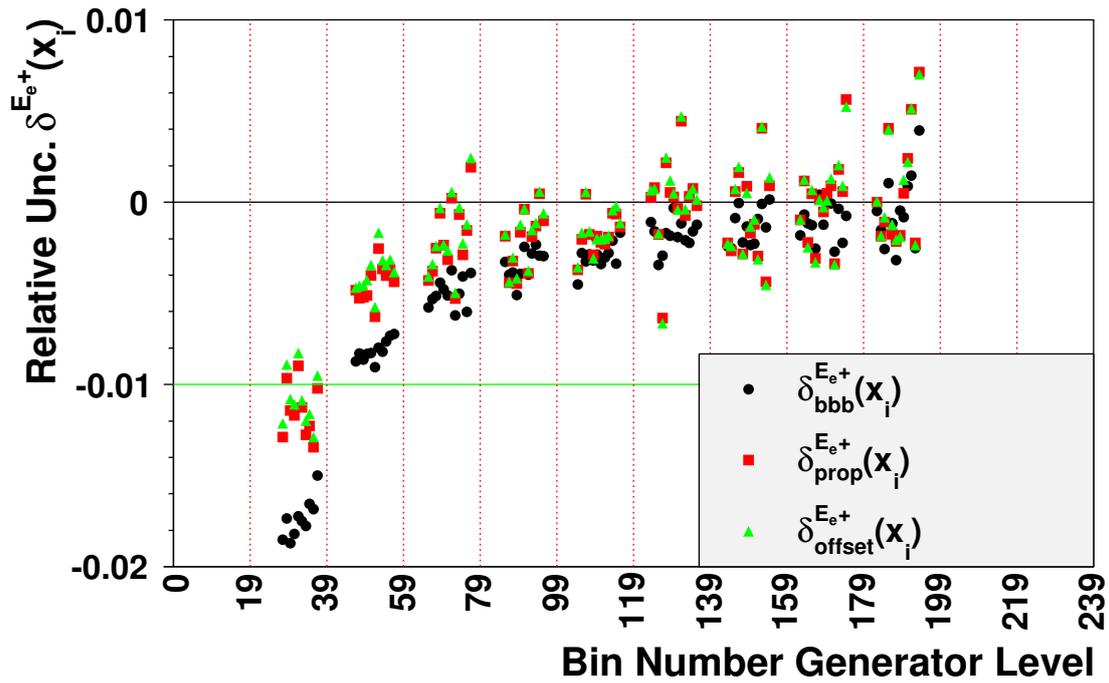
Figure 10.3.10: *Resulting shifts $\delta^{E_e+}(x_i)$ for electron energy scale variation. Input shift: $(\Delta E_e)_+ = +50$ MeV. For the bin numbering scheme, see figure (10.1.1).*
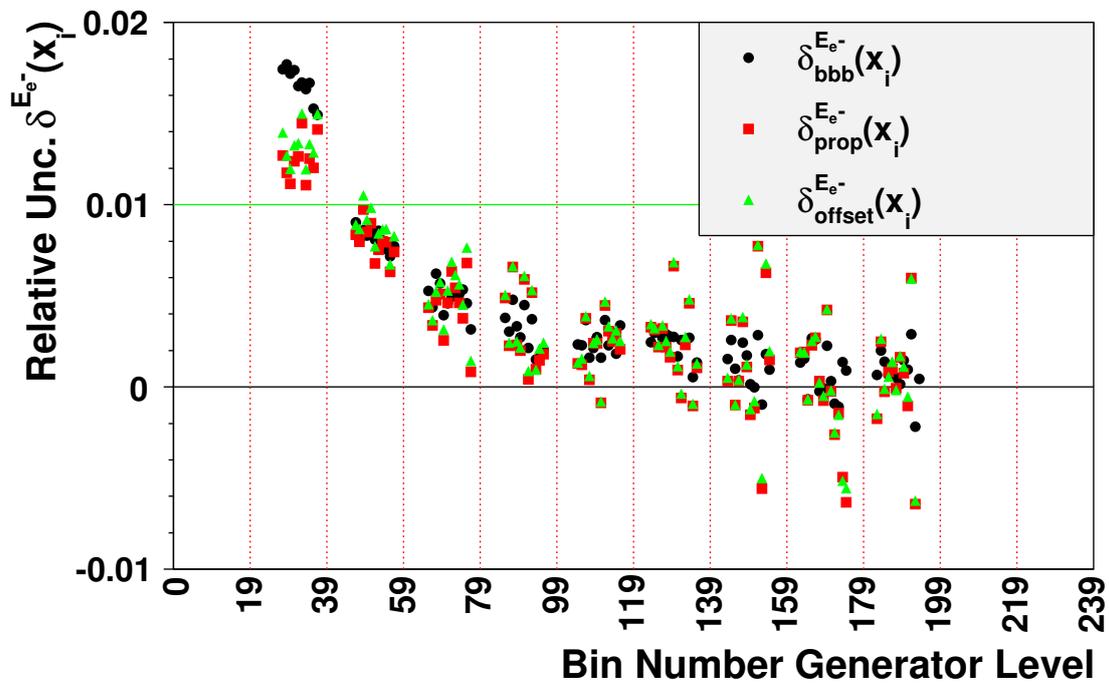
Figure 10.3.11: *Resulting shifts $\delta^{E_e-}(x_i)$ for electron energy scale variation. Input shift: $(\Delta E_e)_- = -50$ MeV. For the bin numbering scheme, see figure (10.1.1).*
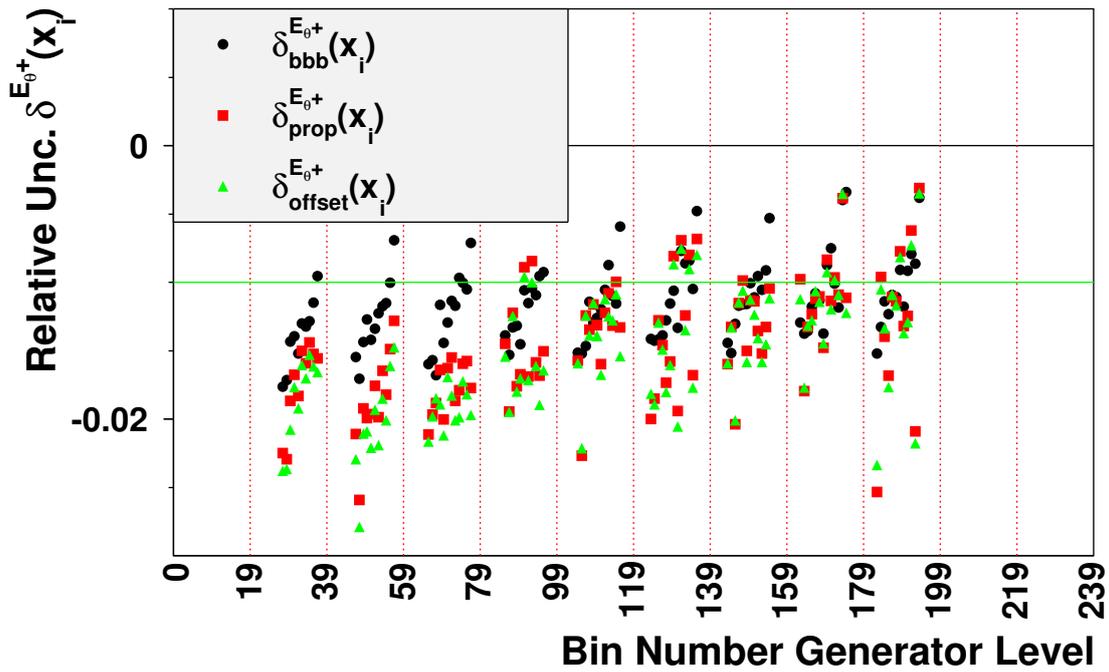
Figure 10.3.12: *Resulting shifts $\delta^{\theta_e+}(x_i)$ for electron energy scale variation. Input shift: $(\Delta\theta_e)_+ = +1$ mrad. For the bin numbering scheme, see figure (10.1.1).*
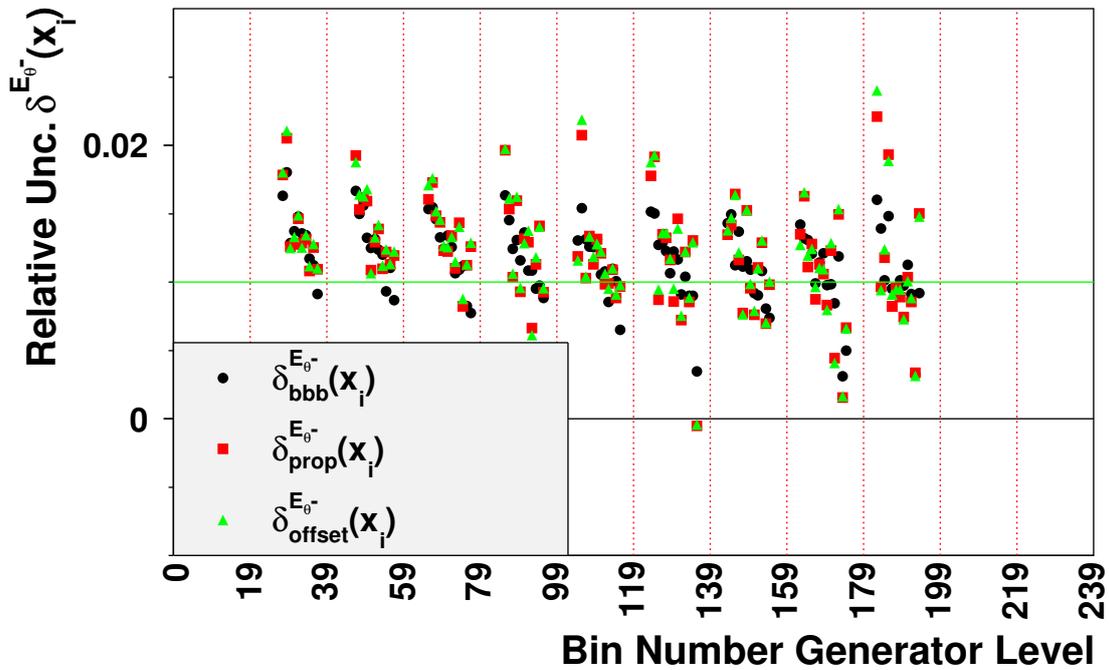
Figure 10.3.13: *Resulting shifts $\delta^{\theta_e-}(x_i)$ for electron energy scale variation. Input shift: $(\Delta\theta_e) = -1$ mrad. For the bin numbering scheme, see figure (10.1.1).*
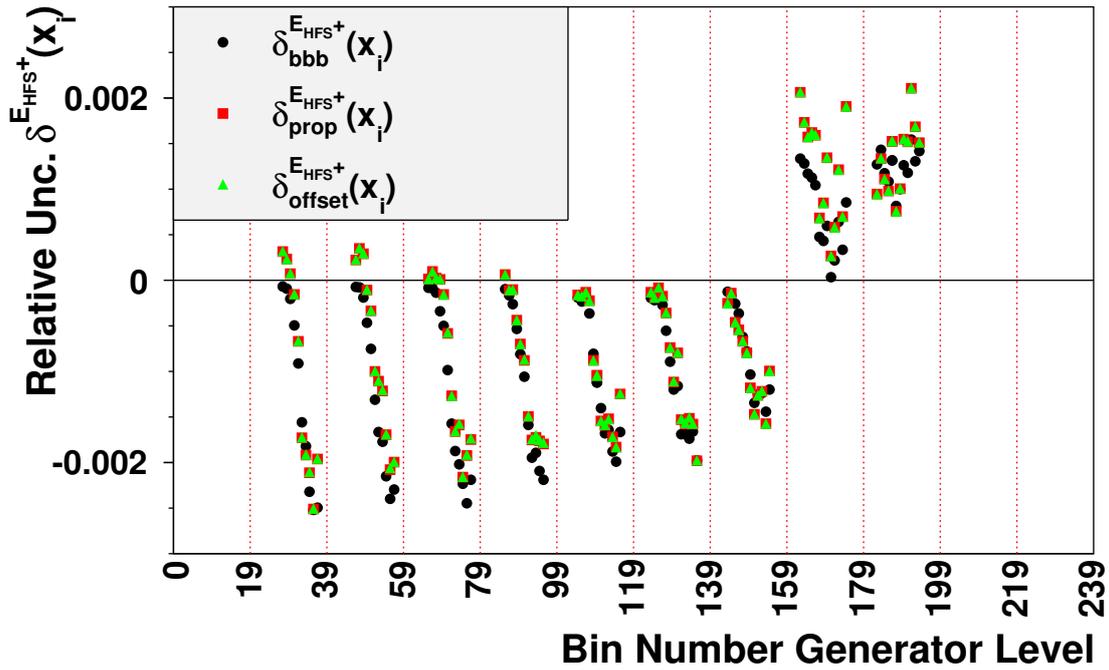
Figure 10.3.14: *Resulting shifts $\delta^{E_{HFS}+}(x_i)$ for hadron energy scale variation. Input shift: $(\Delta E_{HFS})_+ = +4 \% \cdot E_{HFS}$. For the bin numbering scheme, see figure (10.1.1).*
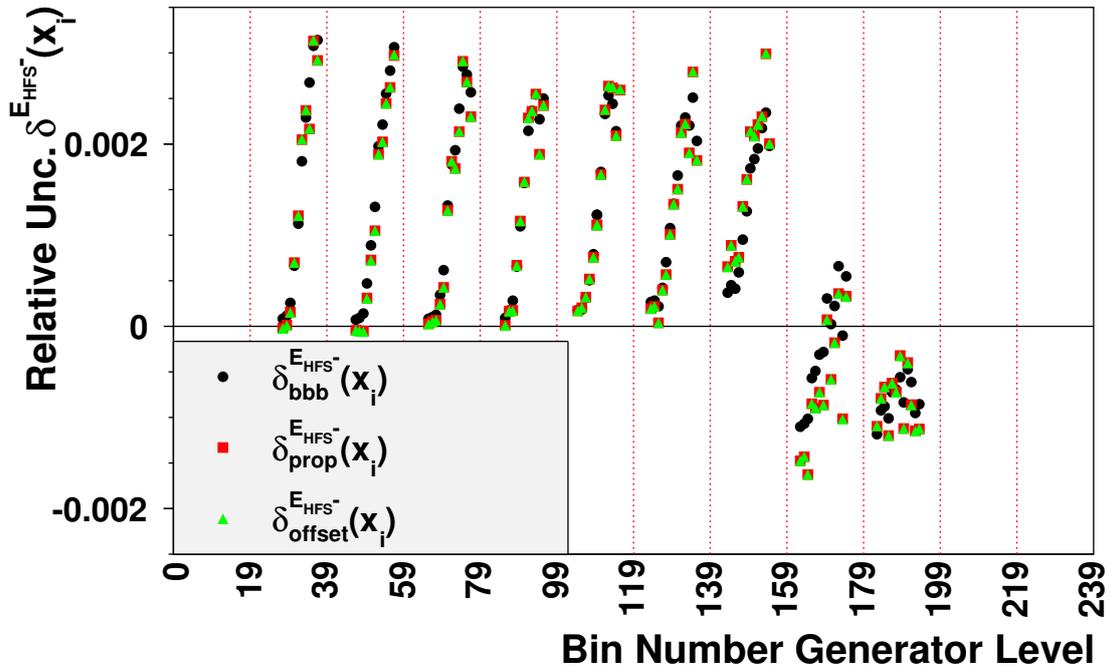
Figure 10.3.15: *Resulting shifts $\delta^{E_{HFS}-}(x_i)$ for hadron energy scale variation. Input shift: $(\Delta E_{HFS})_- = -4 \% \cdot E_{HFS}$. For the bin numbering scheme, see figure (10.1.1).*
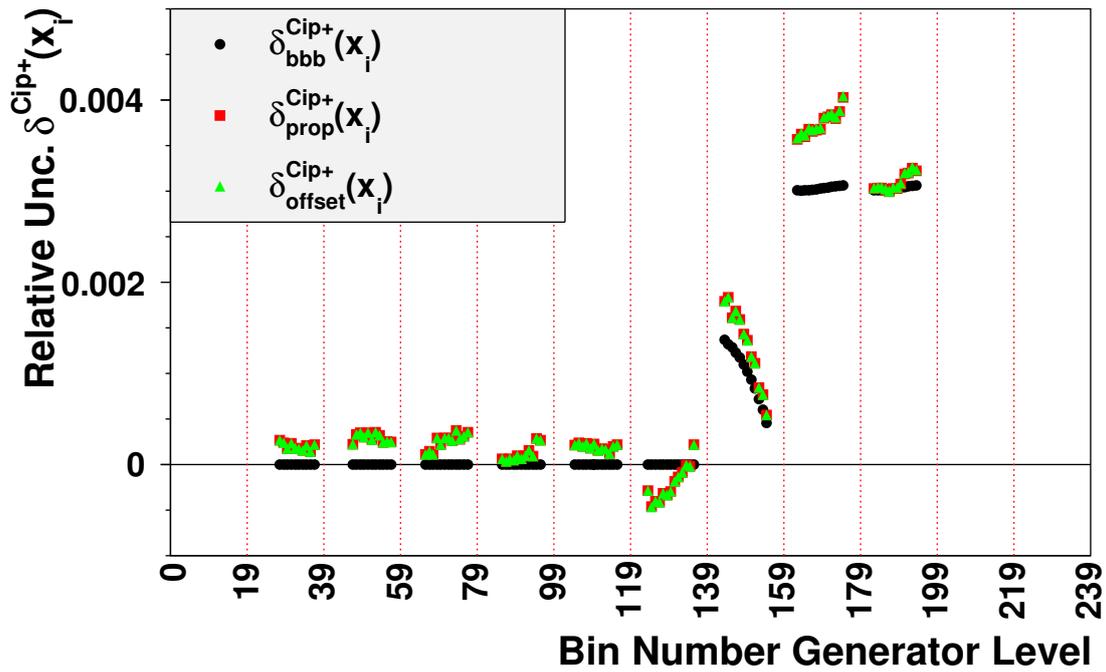
Figure 10.3.16: *Resulting shifts $\delta^{Cip+}(x_i)$ for CIP efficiency variation with shifted weight function w^{Cip+}. See figure (8.2.4) for weight function. For the bin numbering scheme, see figure (10.1.1).*
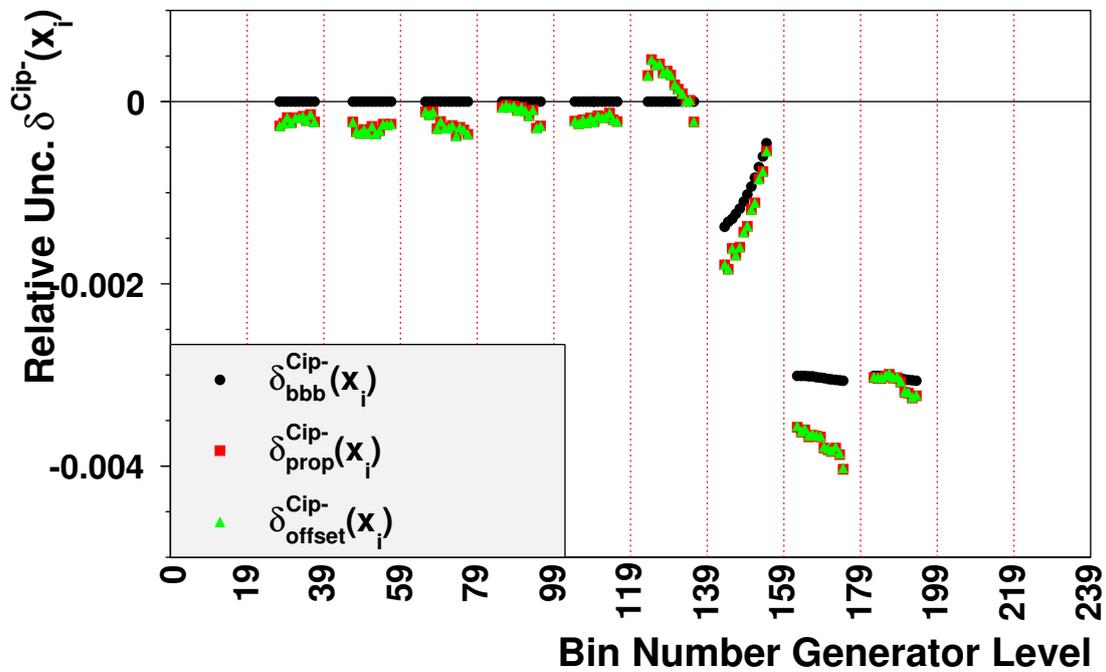
Figure 10.3.17: *Resulting shifts $\delta^{Cip-}(x_i)$ for ECRA cut efficiency variation with shifted weight function w^{Cip-}. See figure (8.2.4) for weight function. For the bin numbering scheme, see figure (10.1.1).*
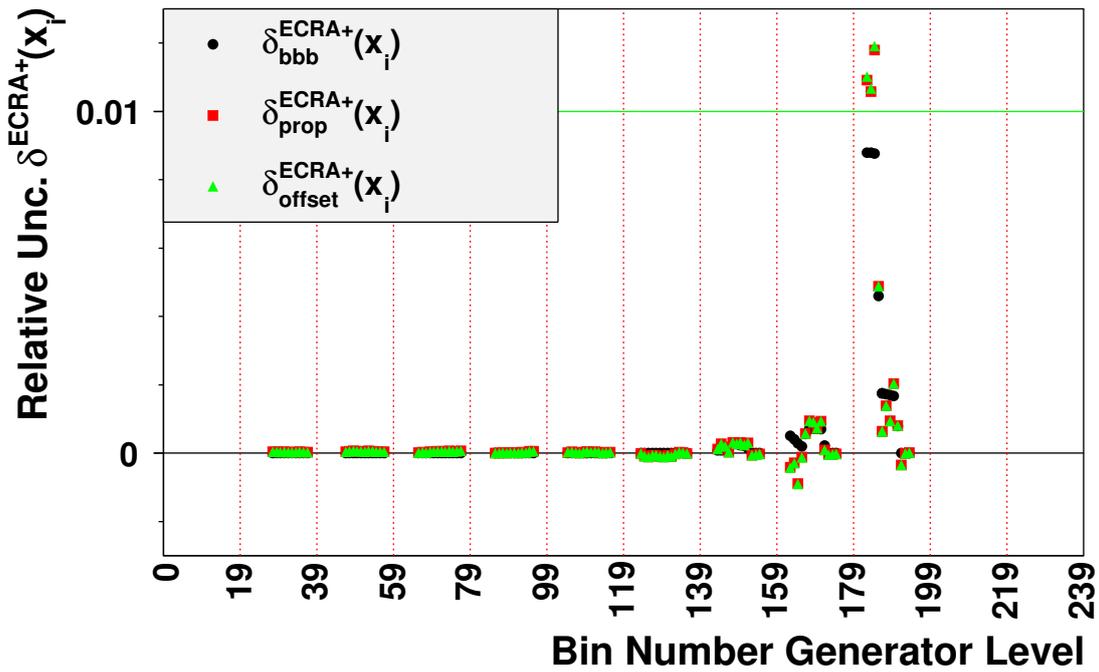
Figure 10.3.18: *Resulting shifts $\delta^{ECRA+}(x_i)$ for CIP efficiency variation with shifted weight function w^{ECRA+}. See figure (8.2.9) for weight function. For the bin numbering scheme, see figure (10.1.1).*
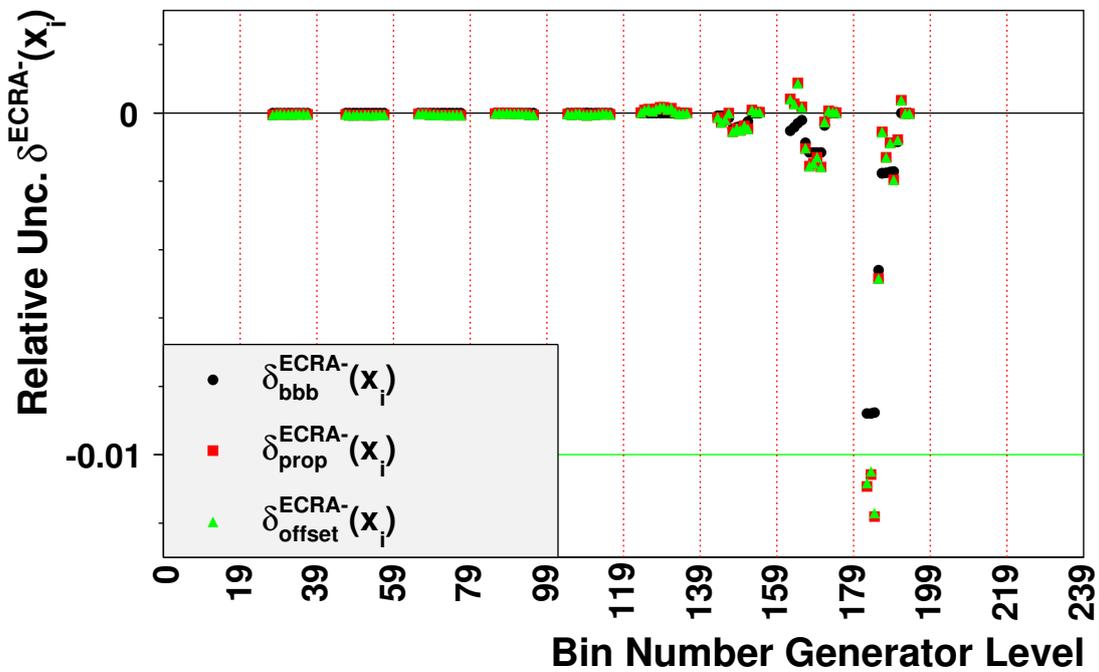
Figure 10.3.19: *Resulting shifts $\delta^{ECRA-}(x_i)$ for CIP efficiency variation with shifted weight function w^{ECRA-}. See figure (8.2.9) for weight function. For the bin numbering scheme, see figure (10.1.1).*
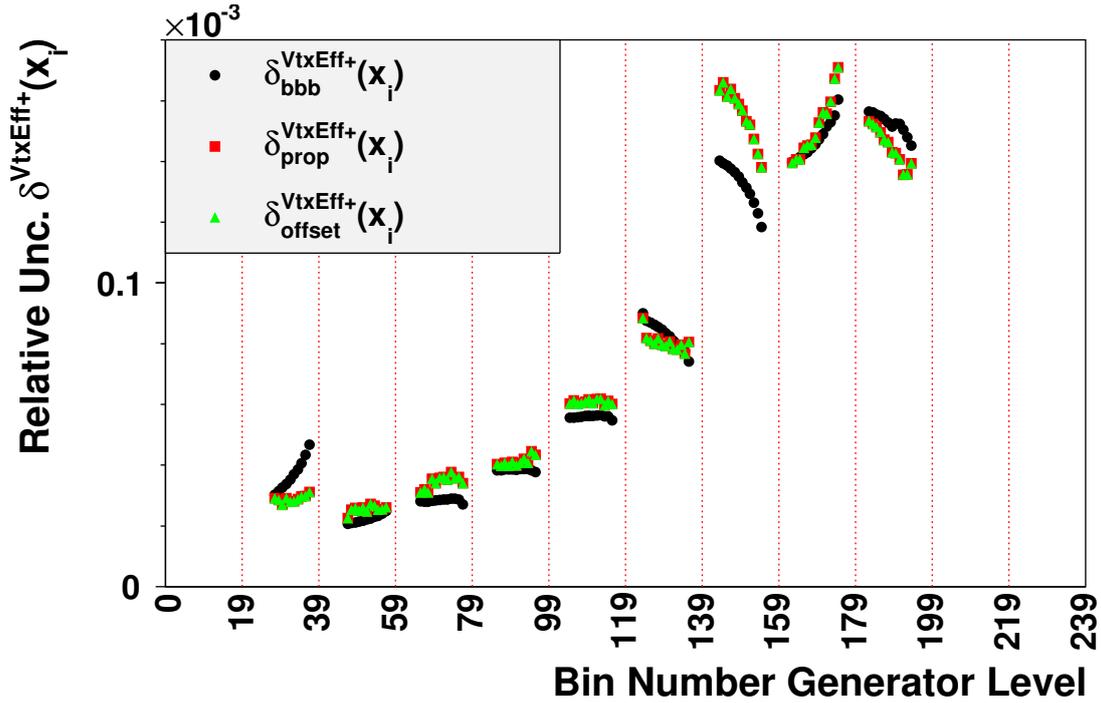
Figure 10.3.20: *Resulting shifts $\delta^{VtxEff+}(x_i)$ for vertex reconstruction efficiency variation with shifted weight function $w^{VtxEff+}$. See figure (8.2.7) for weight function. For the bin numbering scheme, see figure (10.1.1).*
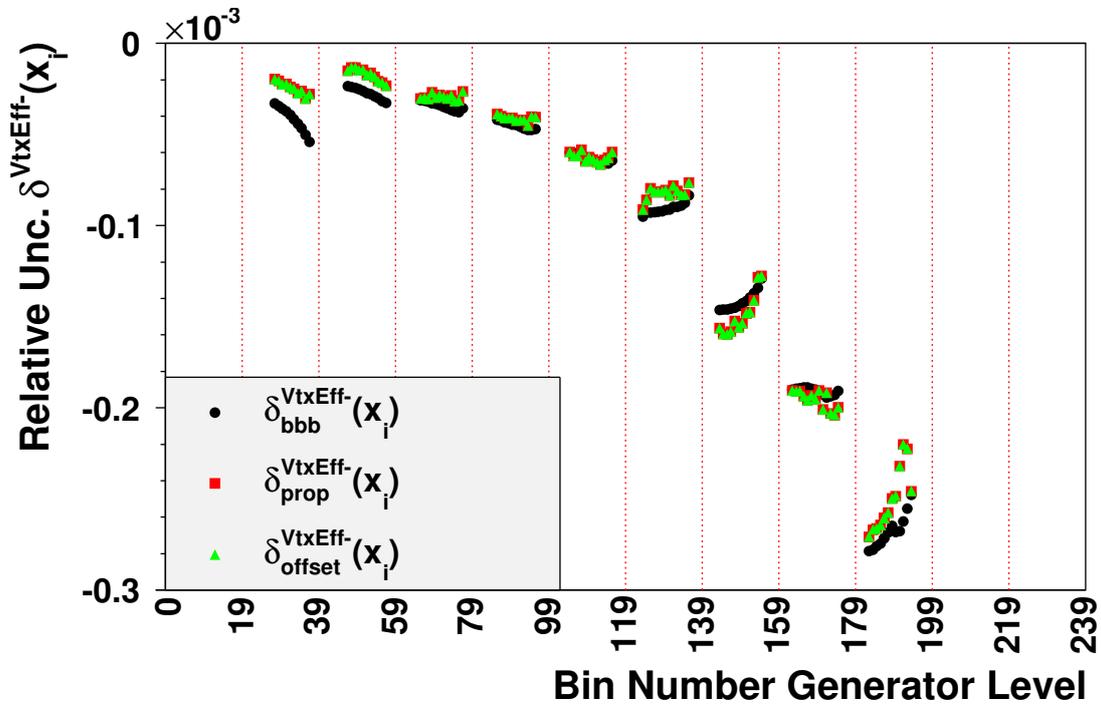
Figure 10.3.21: *Resulting shifts $\delta^{VtxEff-}(x_i)$ for vertex reconstruction efficiency variation with shifted weight function $w^{VtxEff-}$. See figure (8.2.7) for weight function. For the bin numbering scheme, see figure (10.1.1).*
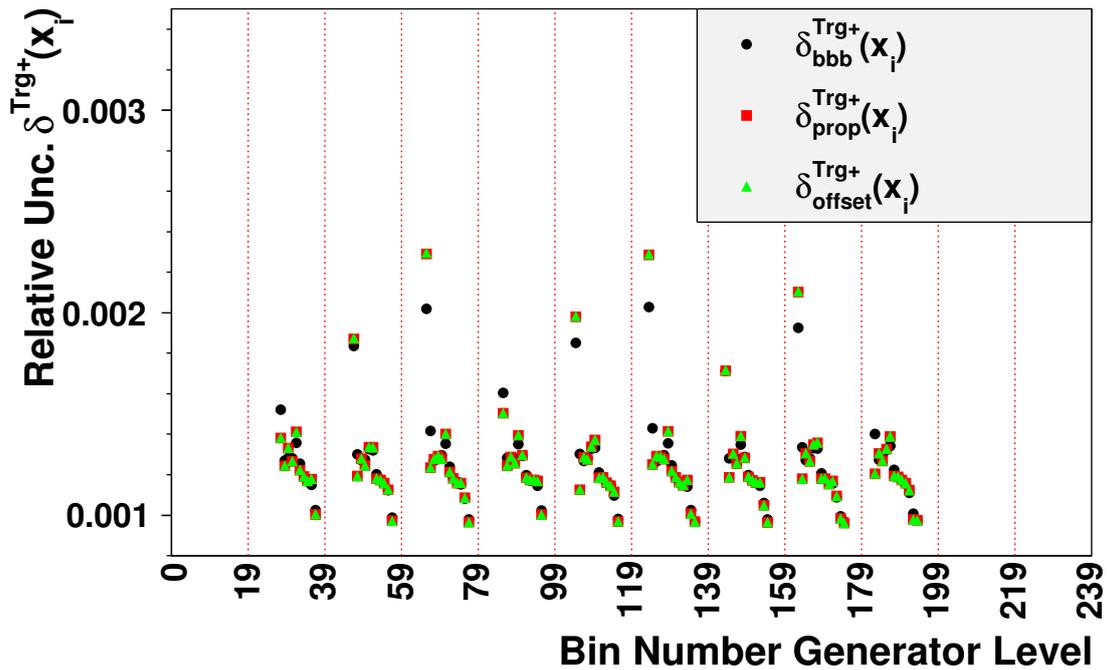
Figure 10.3.22: *Resulting shifts $\delta^{Trg+}(x_i)$ for S3 trigger efficiency variation with shifted weight function w^{Trg+}. See figure (8.2.2) for weight function. For the bin numbering scheme, see figure (10.1.1).*
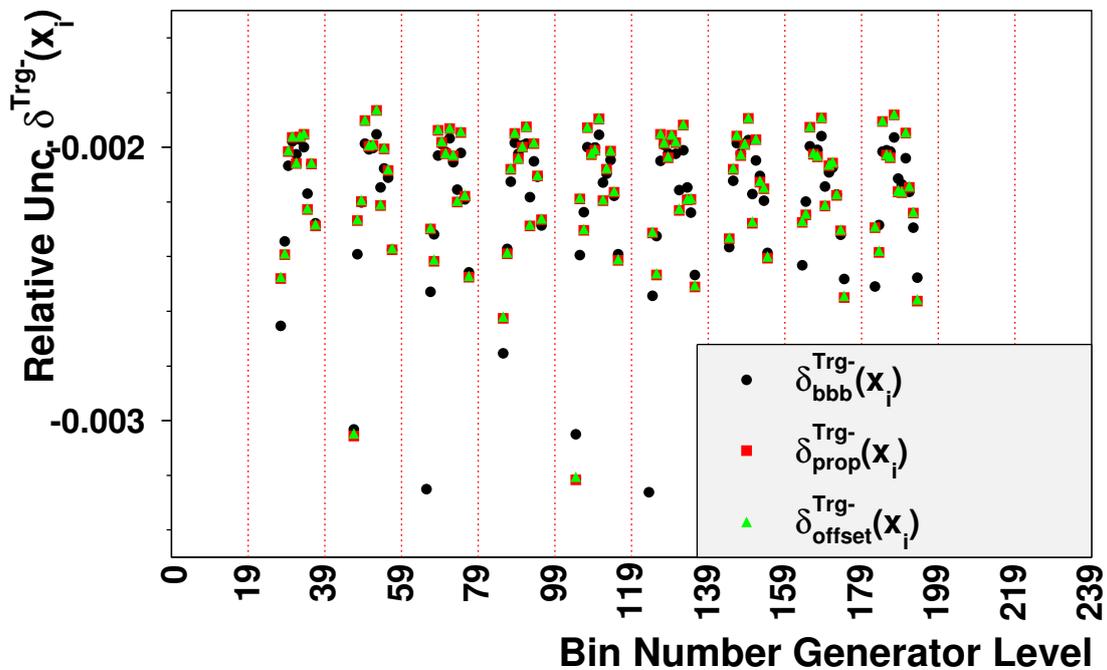
Figure 10.3.23: *Resulting shifts $\delta^{Trg-}(x_i)$ for S3 trigger efficiency variation with shifted weight function w^{Trg-}. See figure (8.2.2) for weight function. For the bin numbering scheme, see figure (10.1.1).*
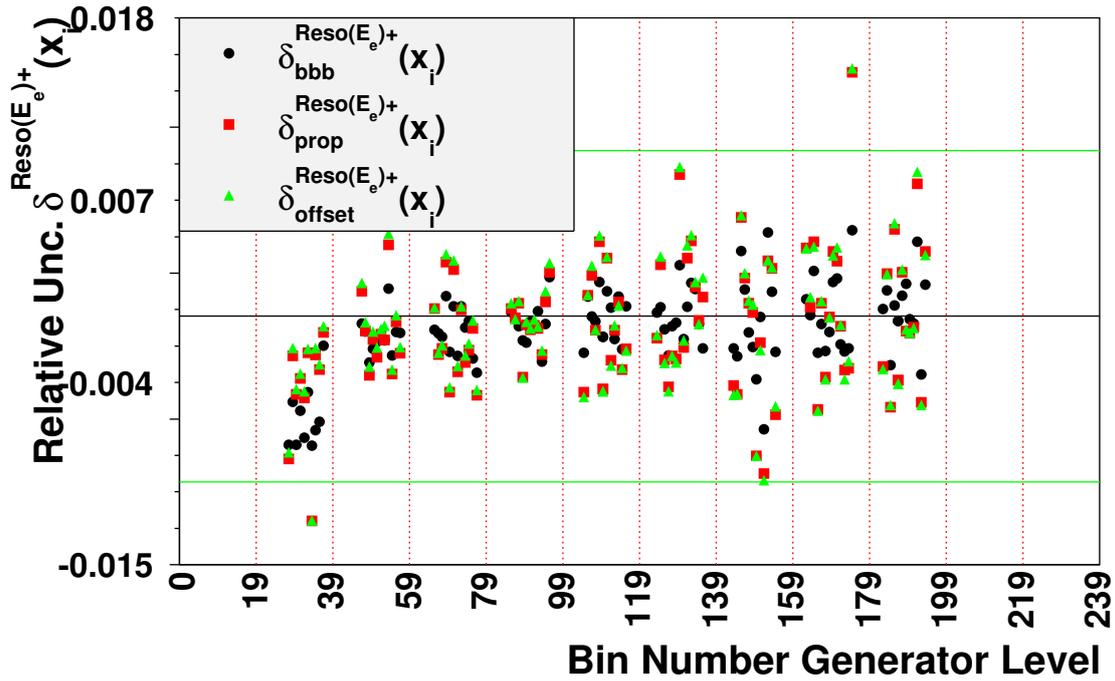
Figure 10.3.24: *Resulting shifts $\delta^{Reso(E_e)}(x_i)$ for energy resolution variation. Effective input shift on resolution: $\Delta E_e = 40$ GeV. For the bin numbering scheme, see figure (10.1.1).*
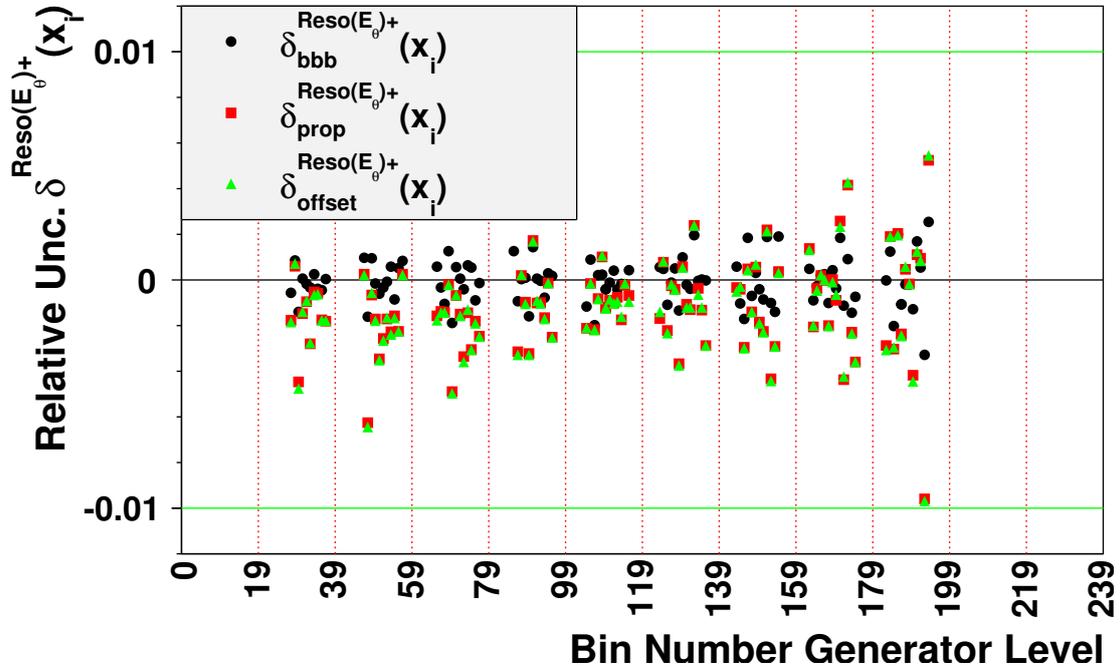
Figure 10.3.25: *Resulting shifts $\delta^{Reso(\theta_e)}(x_i)$ for polar angle resolution variation. Effective input shift on resolution : $\Delta \theta_e = 0.4$ mrad. For the bin numbering scheme, see figure (10.1.1).*
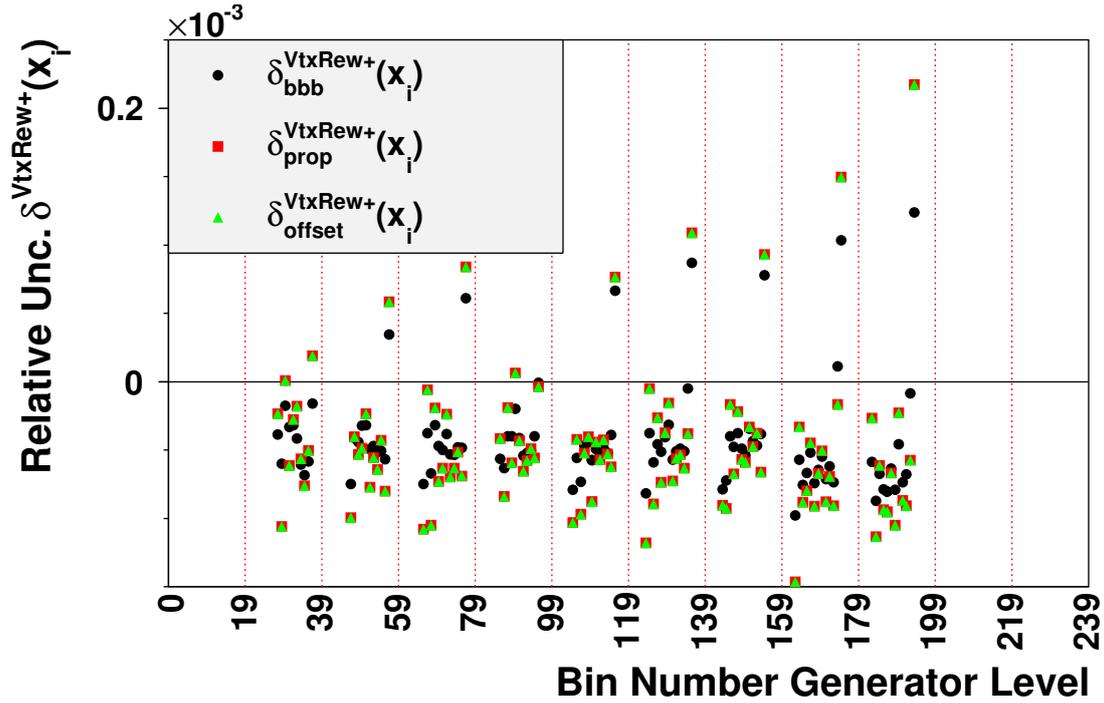
Figure 10.3.26: *Resulting shifts $\sigma^{VtxRew+}(x_i)$ for Z vertex shape reweighting variation with shifted weight function $w^{VtxRew+}$. See figure (8.3.1) for weight function. For the bin numbering scheme, see figure (10.1.1).*
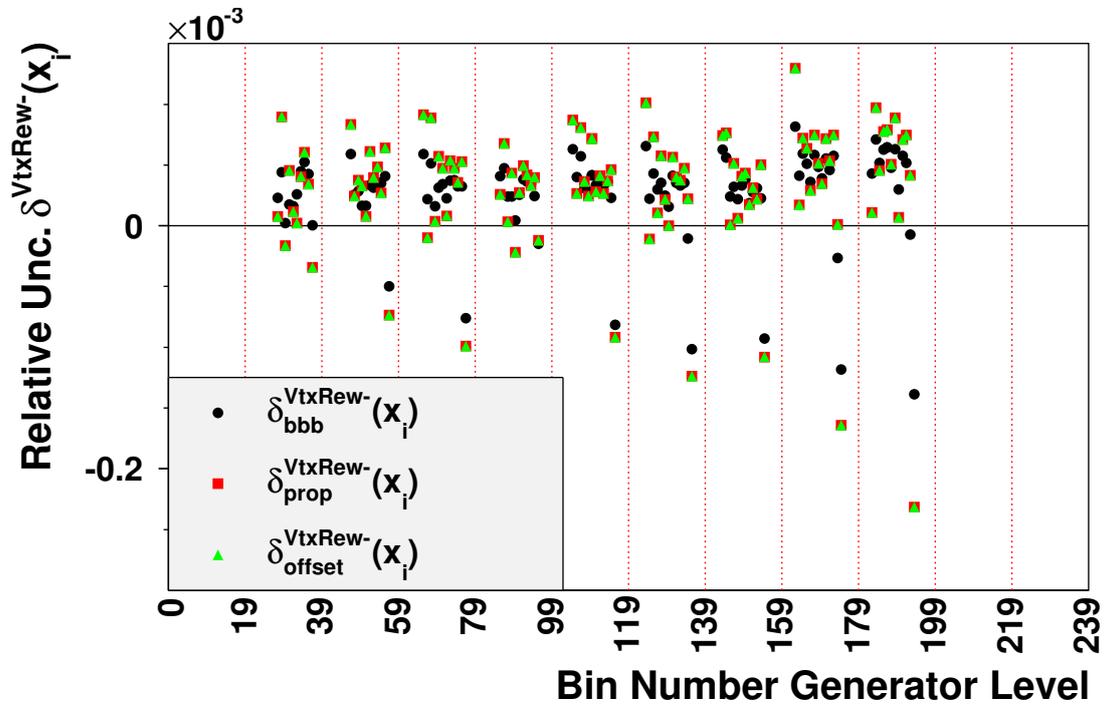
Figure 10.3.27: *Resulting shifts $\sigma^{VtxRew-}(x_i)$ for Z vertex shape reweighting variation with shifted weight function $w^{VtxRew-}$. See figure (8.3.1) for weight function. For the bin numbering scheme, see figure (10.1.1).*
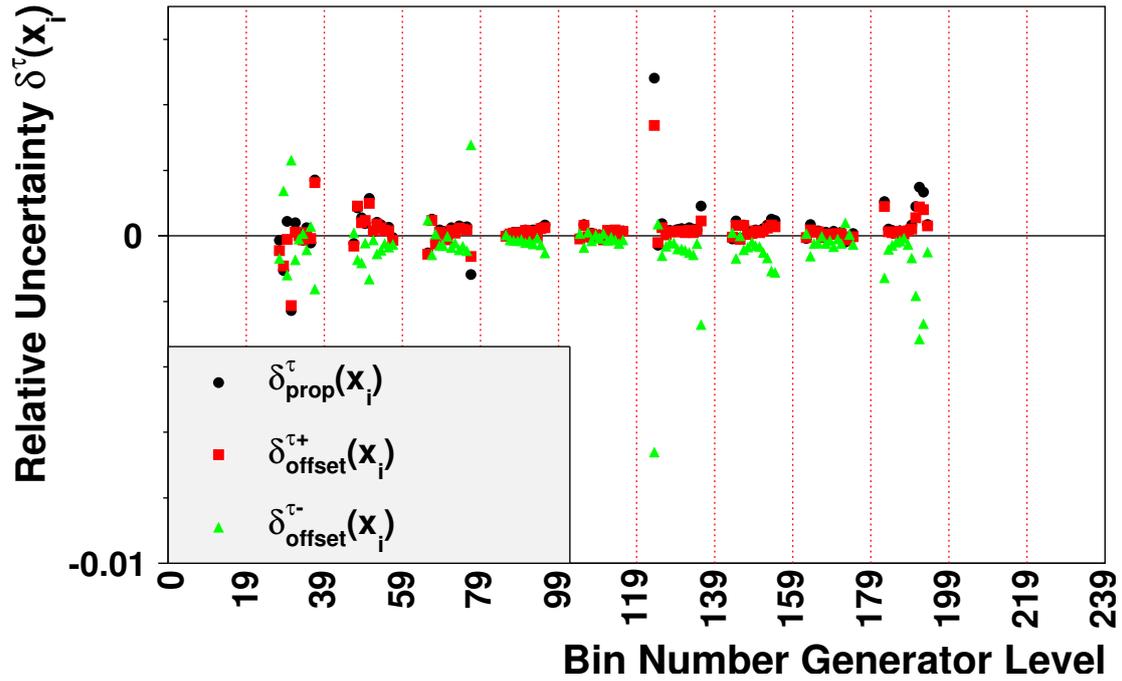
Figure 10.3.28: *Relative shifts $\delta^{\tau+}(x_i)$ and $\delta^{\tau-}(x_i)$ for τ variation with $\Delta\tau \simeq \pm\frac{1}{2}\tau_A$. For the bin numbering scheme, see figure (10.1.1).*
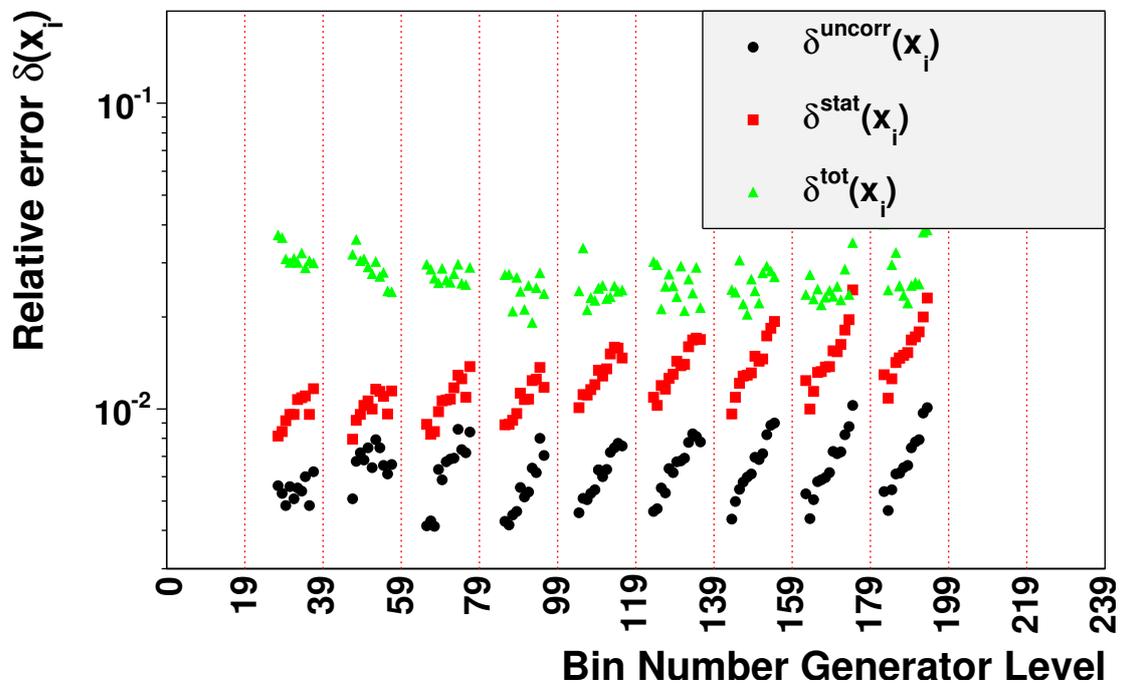
Figure 10.3.29: *Total systematic uncertainties $\delta^{tot}(x_i)$, total uncorrelated systematic uncertainties $\delta^{uncorr}(x_i)$ and total statistical uncertainties $\delta^{stat}(x_i)$. For the bin numbering scheme, see figure (10.1.1).*

10.3.5   Regularization uncertainty

In figure 10.3.28, the impact of the regularization parameter τ on the result is further studied. As input shift, $(\Delta\tau)_{\pm} = \pm\frac{1}{2}\tau_A$ is chosen, i.e. we change the value of τ_A by 50% in both directions. The effect on the unfolding result is studied with the propagation method (4.8.23) and the offset method (4.8.25) according to subsection 4.8.4. The result shows shifts significantly below the permille scale.

Note, that according to the discussion in subsection 4.8.4, these shifts do not enter the total systematic uncertainty.

10.3.6   Total uncertainties

Figure 10.3.29 depicts the total uncorrelated systematic uncertainties $\delta^{uncorr}(x_i)$ according to (4.8.12), the statistical uncertainties $\delta^{stat}(x_i)$ according to (4.8.5) and the total uncertainty $\delta^{tot}(x_i)$ from unfolding, which comprises all statistical, correlated and uncorrelated error sources. Total uncertainties are obtained by quadratic summation.

Chapter 11

Radiative Corrections

The result $\vec{x}$ of the unfolding procedure can be viewed as a measurement, which has sufficiently been corrected for detector effects. Each data point x_i is a direct estimate of the true event count N_i^{true} in bin $\mathcal{B}_i$.

However, as it has been pointed out in chapter 2, the unfolded data distribution $\vec{x}$ contains contributions from a large series of QED processes, such as initial state radiation (ISR), final state radiation (FSR), QED compton events (QEDC) and plain Born level events (BORN). Moreover, virtual vertex corrections contribute to the measured cross section.

In general, the influence of FSR on the measurement is small, because both photon and electron contribute to the electromagnetic energy deposit E_e in the SpaCal. However, the influence of ISR and vertex correction can be substantial. In order to obtain the Born level cross section, bin wise correction factors $\mathcal{R}_i$ are determined and applied to the measurement.

11.1 Determination of Radiative Correction Factors

For the event generation, the `DJANGOH` package has been used. Both simulations have been done using parton density functions from the set `CTEQ6L` [P+02, B+] (LHAPDF code: 10041) and comprise 300 million events each.

For the radiative Monte Carlo, the values $Q^2_{min} = 4$ GeV and $x_{min} = 10^{-5}$ have been used as kinematic boundaries. ISR, FSR, QEDC and Born level events are simulated. Virtual corrections are applied to the QED event vertex. Integration over the phase space gives for the total generator cross section $\sigma_{tot}^{rad} = 347570$ pb or, equivalently:

$$\mathcal{L}^{rad} = 863.1 \text{ pb}^{-1} \tag{11.1.1}$$

For the non-radiative Monte Carlo, the values $Q^2_{min} = 4$ GeV and $x_{min} = 10^{-4}$ have been used as kinematic boundaries. Only Born level events are simulated. No virtual corrections are applied. Integration over the phase space gives for the total generator cross section $\sigma_{tot}^{norad} = 293340$ pb or, equivalently:

$$\mathcal{L}^{norad} = 1022.7 \text{ pb}^{-1} \tag{11.1.2}$$

In figure 11.1.1 control plots for the generator variables x_e, y_e and Q_e^2 are drawn. Note, that these quantities explicitly refer to the electron kinematics according to section 2.1. Clearly visible is the effect of initial state radiation (ISR). This process effectively reduces the available center of mass energy s, leading to smaller Q_e^2 and higher y_e, see (3.5.2) and (3.5.1).

In figure 11.1.2 the radiative correction factors $\mathcal{R}_i$ are depicted for each bin $\mathcal{B}_i$. These typically take on values around $\sim 80\,\%$, while exhibiting a falling y-dependence.
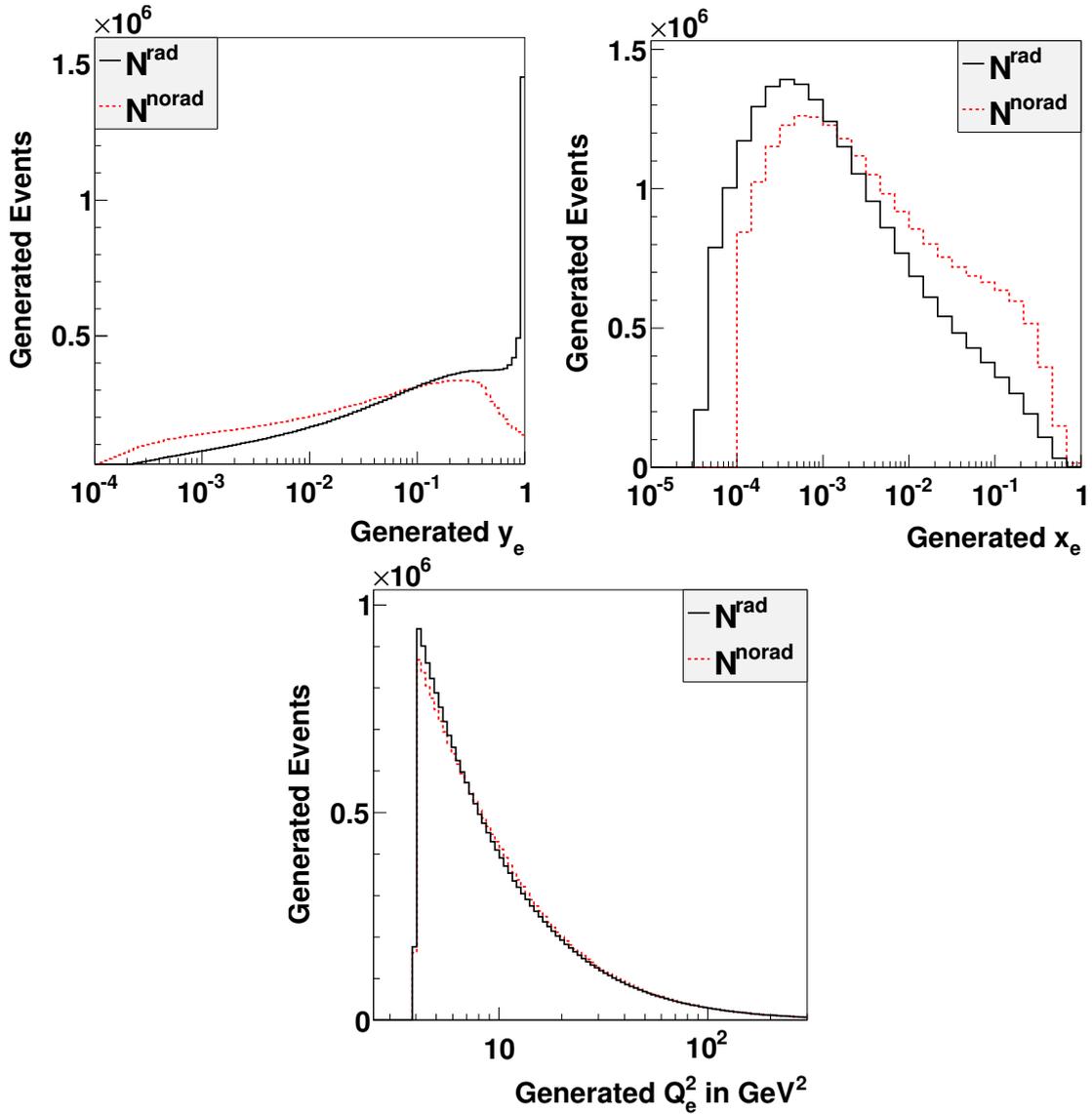
Figure 11.1.1: *Generator level distributions with (N^{rad}) and without (N^{norad}) radiative corrections. Top Left: Inelasticity y_e. Top right: Bjørken variable x_e. Bottom: Squared momentum exchange Q_e^2.*
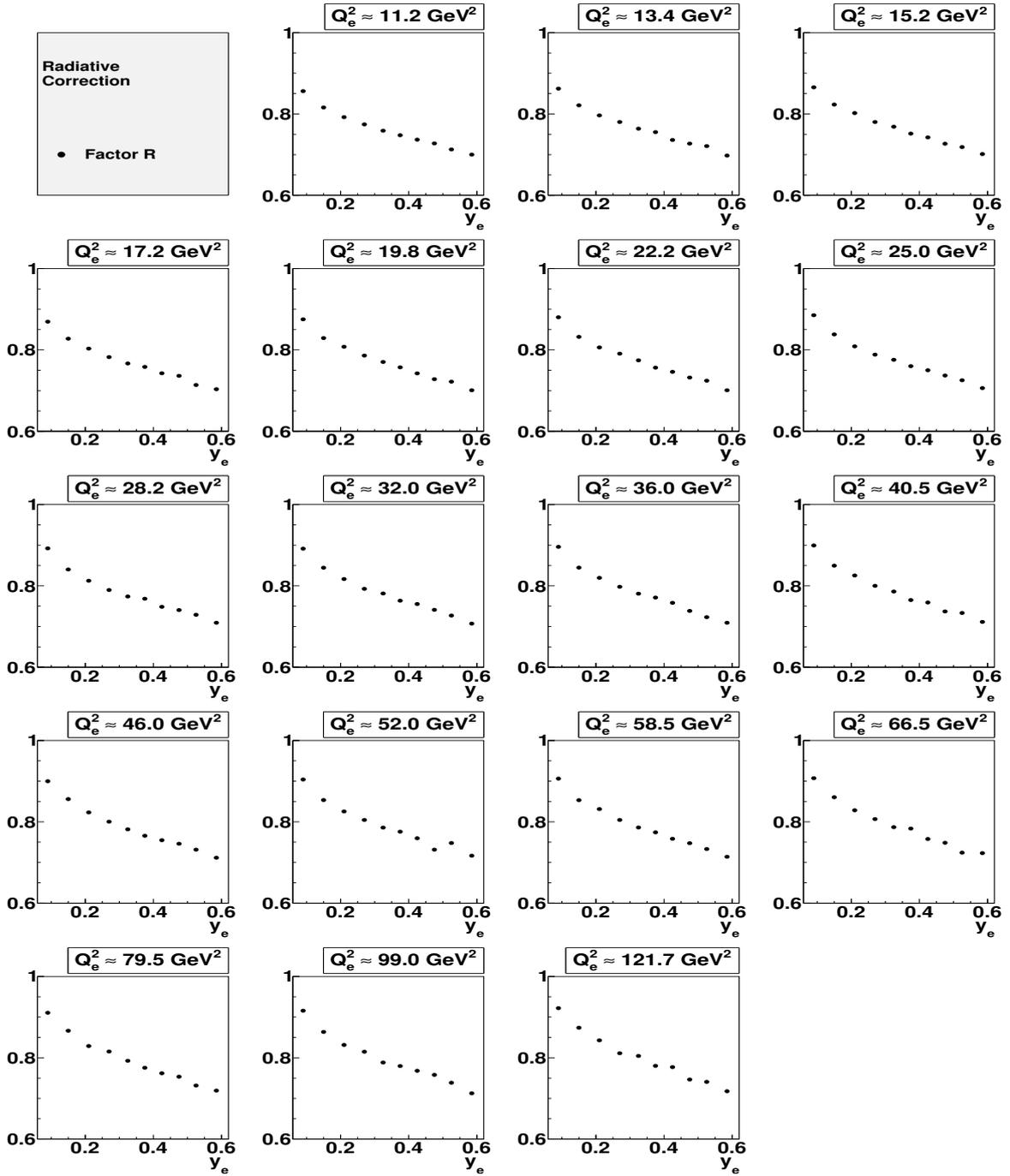
Figure 11.1.2: *Correction factors $\mathcal{R}_i$ for different bins in Q_e^2.*

11.2 Systematic Uncertainties from Radiative Corrections

For the systematic uncertainty which is introduced by the radiative correction factors $\mathcal{R}_i$ on the Born level cross section σ_r^{Born}, a global value of

$$\delta^{rad}\left(\mathcal{R}_i\right) \simeq 0.3\% \tag{11.2.1}$$

is adopted. In [A$^+$09], this is number is obtained from comparisons of the analytical cross section computations employing different mathematical approaches. In particular, the calculations from the packages HERACLES [KSM92] and HECTOR [ABB$^+$96] were compared, which feature a leading order calculation and a leading-logarithm approach, repetively.

Chapter 12

Cross Sections and F_2 Structure Function

In this chapter, the result of the measurement is discussed. We present both the reduced cross section σ_r and the F_2 structure function and compare it to previous measurements. Moreover, we discuss differences between the cross sections from the unfolding method and the bin-by-bin method.

12.1   Reduced Cross Sections

In figure 12.3.1, the reduced Born level cross section σ_r^{Born} according to (2.2.2) and (5.1.9) is drawn. It is contrasted to the functional shape according to the pdf set H1PDF2000 ($\sigma_r^{H1PDF2000}$) [A$^+$03] . The error bars depict the statistical uncertainty $\Delta^{stat}\left(\sigma_r^{Born}\right)$, see (4.8.5), and the total uncertainty $\Delta^{tot+r}\left(\sigma_r^{Born}\right)$ including the contributions of the radiative corrections, see (5.2.1). An additional luminosity error of $\delta\left(\mathcal{L}_{data}\right) = 2\%$ needs to be added to obtain the full uncertainty, see section 6.1. Note, that the bin center correction is done implicitly by assigning each data point a bin center (Q_e^2, y_e) according to the mean value theorem. See the discussion in section 5.3 for details.

We continue the comparison in figure 12.3.2, which shows the ratio of σ_r^{Born} and $\sigma_r^{H1PDF2000}$. The error bars depict the relative statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ and the total uncertainty $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ of σ_r^{Born}. While the overall ratio is close to 1, a rising trend is visible with both y_e and Q_e^2. Note the remaining statistical fluctuations, which arise due to the rather conservative choice of the regularization level ($\log\tau = \log\tau_A \simeq -4.29$, see chapter 10). These fluctuations can be accounted for in a fit procedure by the full correlation matrix $\boldsymbol{Corr}_{\vec{\sigma}}^{stat+uncorr}$, see figure 12.3.3. It provides correlations for both statistical and uncorrelated systematic error sources.

In figure 12.3.4, both the reduced Born level cross section σ_r^{Born} and the reduced radiative cross section σ_r^{rad} are given. They are compared to the functional shape according to the data set H1PDF2009 ($\sigma_r^{H1PDF2009}$) [Kre, A$^+$09] and the generator cross section of the full radiative Monte Carlo simulation (σ_r^{gen}), which is based on H1PDF2009 by reweighting. The error bars depict the total uncertainties $\Delta^{tot}\left(\sigma_r^{rad}\right)$ and $\Delta^{tot+r}\left(\sigma_r^{Born}\right)$, while for the Born level cross section the contributions of the

radiative corrections are included. The luminosity error is not included.

The comparison is continued in figure 12.3.5, which shows the ratio of σ_r^{Born} and $\sigma_r^{H1PDF2009}$. Again, the error bars depict the relative statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ and the total uncertainty $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ of σ_r^{Born}. In addition to the slope with y_e and Q_e^2, which reoccurs in this comparison, a global difference of roughly 2-3 % is visible between our measurement σ_r^{Born} and $\sigma_r^{H1PDF09}$. Therefore, the measurement of this analysis seems to affirm rather the `H1PDF2000` fit than the one from `H1PDF2009`. Note however, that the luminosity measurement is afflicted with the rather large systematic uncertainty of $\delta\left(\mathcal{L}_{data}\right) \simeq 2\%$, see (6.1.2).

The total uncertainties are depicted in figure 12.3.6. The points shows the statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ and the cumulated error $\delta^{corr}\left(\sigma_r^{Born}\right)$ of all correlated systematic error sources. The statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ rises with both y_e and Q_e. The correlated error sources are dominated by the scale uncertainties on E_e and θ_e. While the energy scale uncertainty mainly affects the low y_e region (up to 2% uncertainty), the polar angle uncertainty affects the region of low Q_e^2 (small scattering angles, up to 2% uncertainty). At very low E_e and R_e the systematic uncertainty on the ECRA cut efficiency correction comes into play (up to 1.2%), which is visible at very high bins in y_e.

The lines in 12.3.6 depict cumulated errors. $\delta^{stat+uncorr}\left(\sigma_r^{Born}\right)$ comprises the statistical error and the total uncertainty of all uncorrelated error sources, which originate from the Monte Carlo statistics. The total error $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ also includes the correlated error sources and the uncertainty from the radiative corrections. Finally, $\delta^{tot+r+l}\left(\sigma_r^{Born}\right)$ contains all error sources including the luminosity error, see (5.2.2). For $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ we obtain a total precision of roughly 2.5 % troughout the kinematical plane.

In appendix A, the measured cross sections, its uncertainties and global correlations are given in a tabular form.

12.2　Comparison with Bin-By-Bin Method

In figure 12.3.7 the result σ_r^{Born} is compared to the bin-by-bin method $\sigma_r^{Born,bbb}$. The error bars correspond to the relative statistical uncertainties $\delta^{stat}\left(\sigma_r^{Born}\right)$ of the unfolded result. The high level of compatability of both results σ_r^{Born} and $\sigma_r^{Born,bbb}$ is evident, the ratio is close to 1 throughout the covered phase space.

As expected, the ratio of statistical uncertainties, $\delta^{stat}\left(\sigma_r^{Born}\right)/\delta^{stat}\left(\sigma_r^{Born,bbb}\right)$ is significantly greater than 1, see figure 12.3.8. As discussed in chapter 10, this reflects both the bin-to-bin correlations as well as the underestimation of statistical uncertainty by the bin-by-bin method.

12.3　Structure Function F_2

As discussed in chapter 2, the structure function $F_2\left(x, Q^2\right)$ can be determined from the reduced cross $\sigma_r^{Born}\left(x, Q^2\right)$ by the application of a small correction which accounts for $F_L\left(x, Q^2\right)$, see (2.2.1). The result is shown in figure 12.3.9. For $F_L\left(x, Q^2\right)$, the

pdf set `H1PDF2000` has been employed. Again, the error bars depict the statistical uncertainty $\Delta^{stat}(F_2)$ and the total uncertainty $\Delta^{tot+r}(F_2)$.

For comparison, the structure functions according to `H1PDF2000` [A$^+$03] and `H1PDF2009` [Kre, A$^+$09] are given ($F_2^{H1PDF2000}$ and $F_2^{H1PDF2009}$), together with a fit based on ZEUS data ($F_2^{ZEUS2005}$) [C$^+$03].
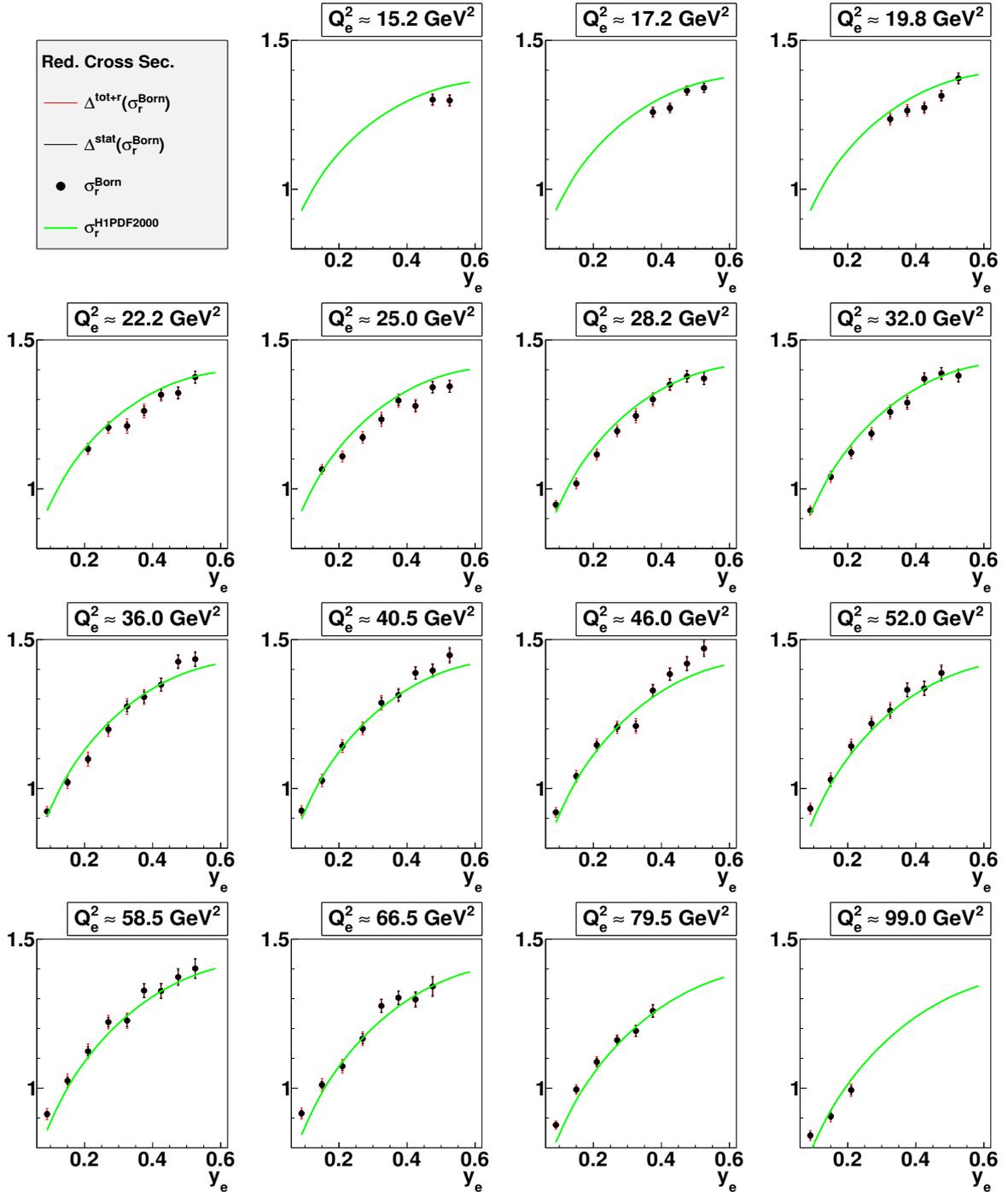
Figure 12.3.1: *The measured reduced cross section* σ_r^{Born} *is shown. The error bars correspond to the statistical uncertainty* $\Delta^{stat}\left(\sigma_r^{Born}\right)$ *and the total uncertainty* $\Delta^{tot+r}\left(\sigma_r^{Born}\right)$ *including the contributions of the radiative corrections, see (5.2.1). An additional luminosity error of* $\delta\left(\mathcal{L}_{data}\right) = 2\%$ *needs to be added to obtain the full uncertainty. For comparison, the functional shape according to the pdf set* H1PDF2009 $\left(\sigma_r^{H1PDF09}\right)$ *[Kre, A$^+$09] is plotted.*
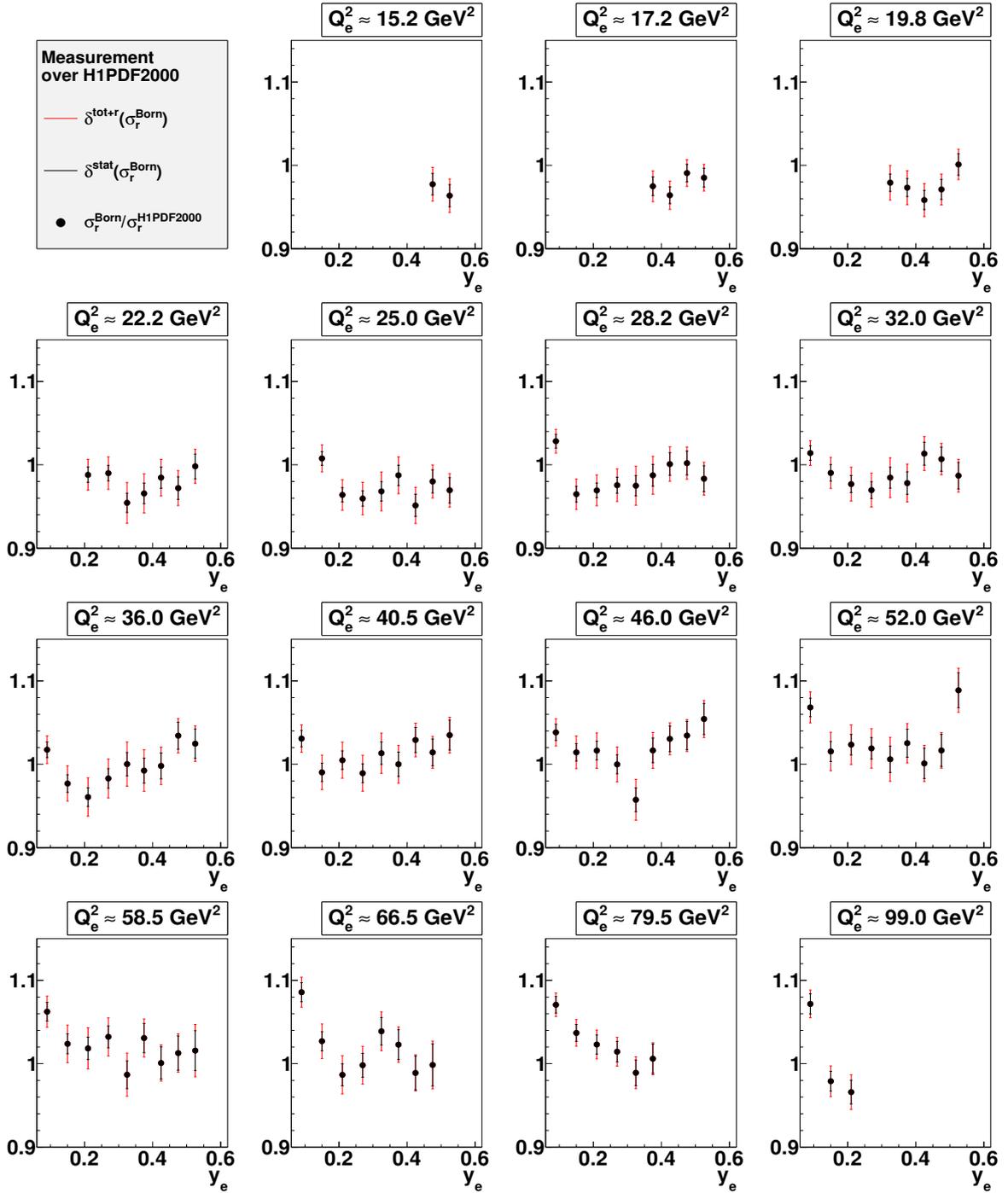
Figure 12.3.2: *Ratio of σ_r^{Born} and $\sigma_r^{H1PDF2000}$ (*`H1PDF2000`*) [A+03]. The error bars correspond to the statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ and the total uncertainty $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ including the contributions of the radiative corrections, see (5.2.1). An additional luminosity error of $\delta\left(\mathcal{L}_{data}\right) = 2\%$ needs to be added to obtain the full uncertainty.*
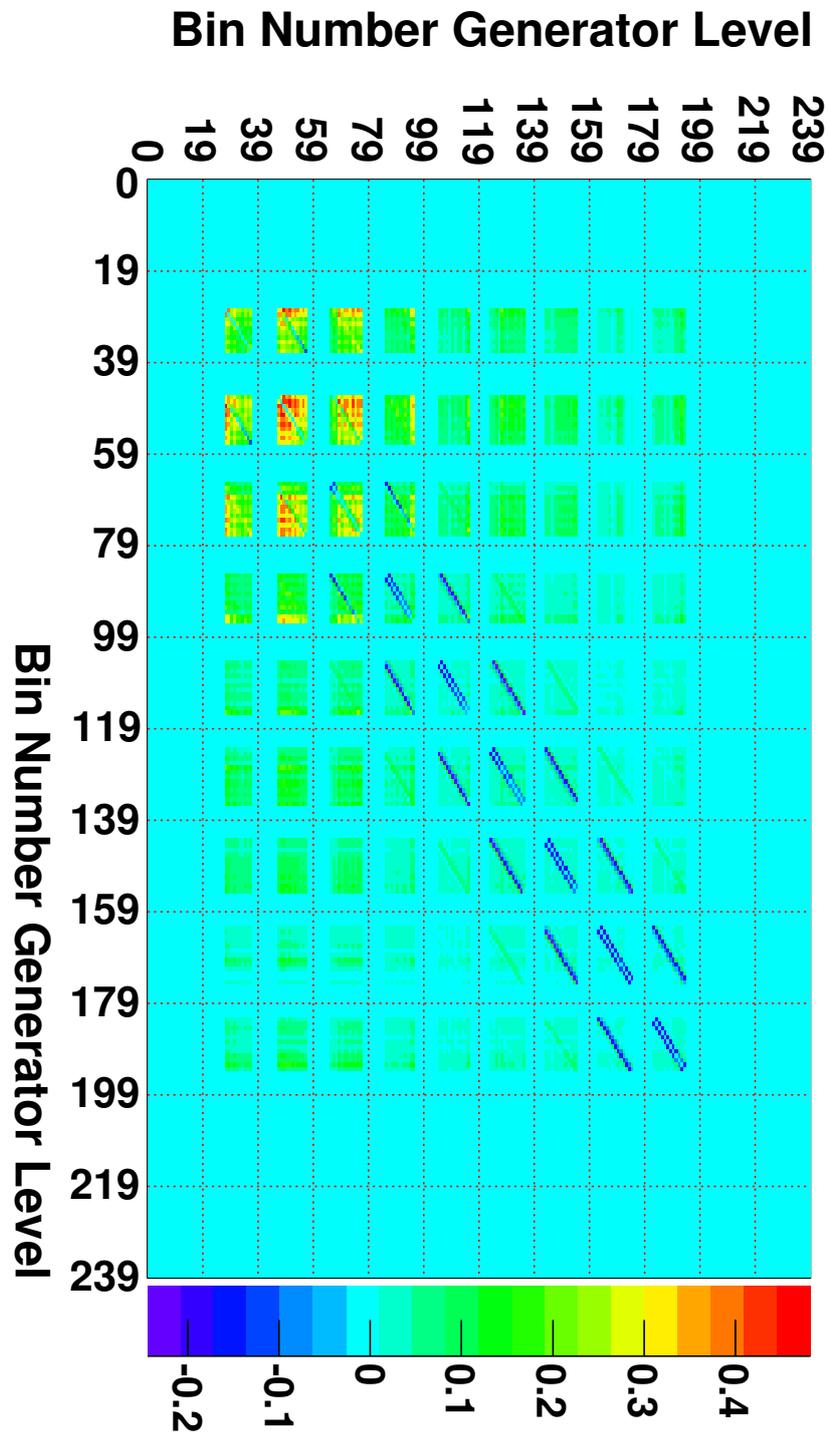
Figure 12.3.3: *Correlation matrix $\boldsymbol{Corr}^{stat+uncorr}_{\vec{\sigma}}$ accounting for statistical bin-to-bin correlations as well as correlations from uncorrelated error sources.*
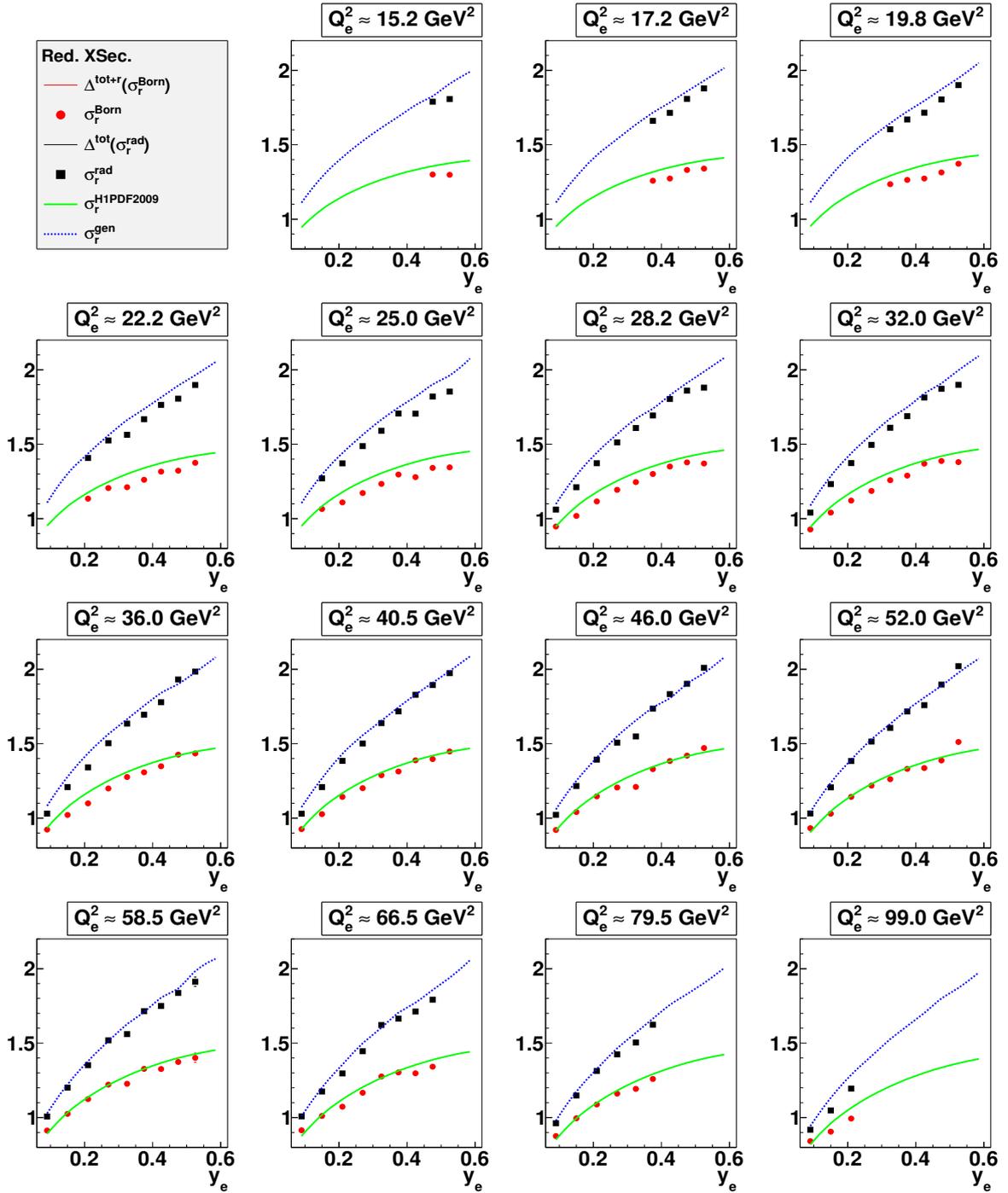
Figure 12.3.4: *The measured reduced cross sections. Both, σ_r^{Born} for the Born level result and σ_r^{rad} for the full radiative measurement. The error bars depict the total uncertainties $\Delta^{tot}\left(\sigma_r^{rad}\right)$ and $\Delta^{tot+r}\left(\sigma_r^{Born}\right)$, while for the Born level cross section the contributions of the radiative corrections are included. An additional luminosity error of $\delta\left(\mathcal{L}_{data}\right) = 2\%$ needs to be added to obtain the full uncertainty. For comparison, the functional shape according to the pdf set H1PDF2009 ($\sigma_r^{H1PDF09}$) [Kre, A+09] is plotted, as well as the generator cross section σ_r^{gen} of the full radiative Monte Carlo model (reweighted to H1PDF2009). The error bars correspond to the statistical uncertainty $\Delta^{stat}\left(\sigma_r^{Born}\right)$ and $\Delta^{stat}\left(\sigma_r\right)$, respectively.*
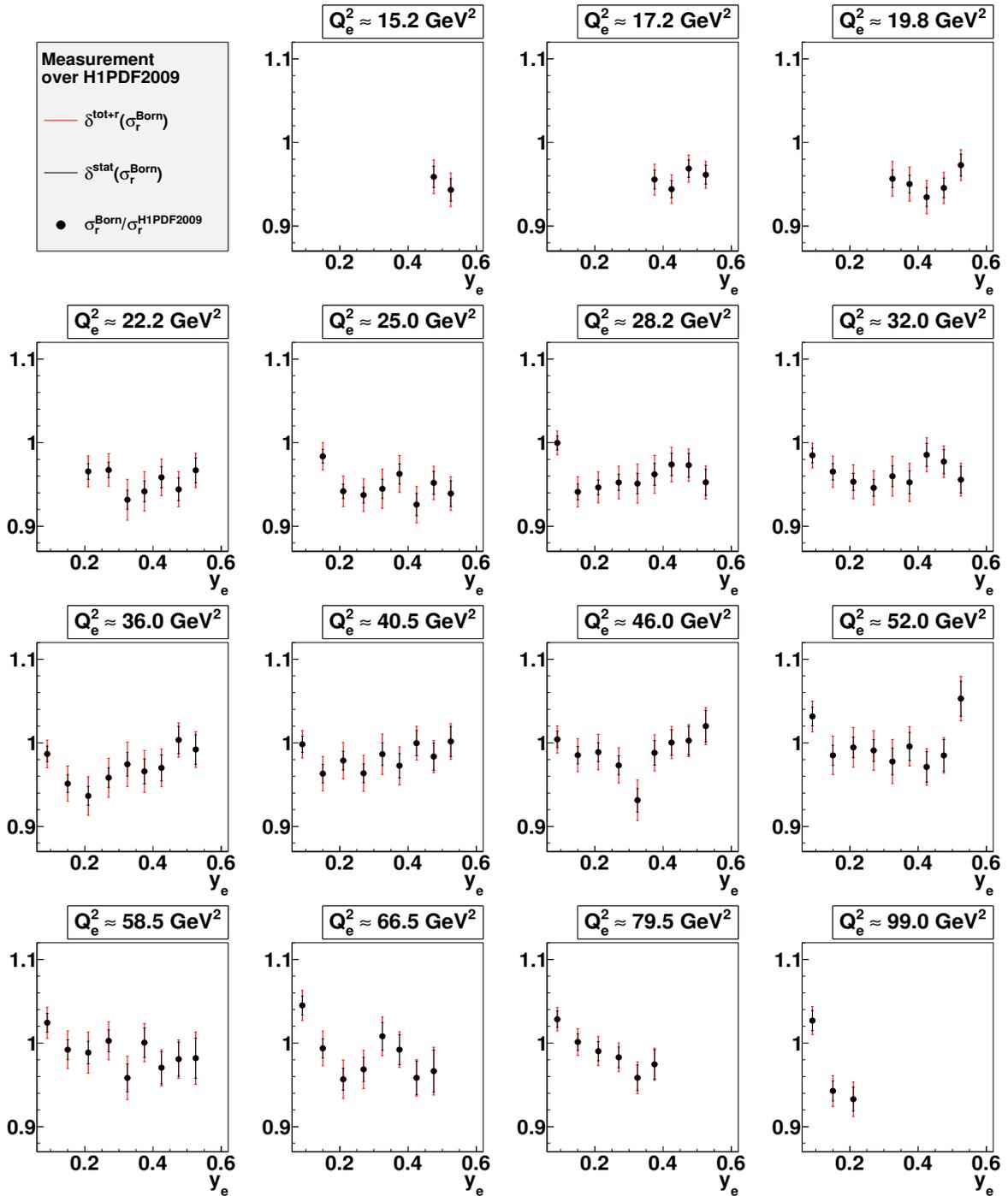
Figure 12.3.5: *Ratio of σ_r^{Born} and $\sigma_r^{H1PDF2009}$ (*H1PDF2009*) [Kre, A$^+$09]. The error bars correspond to the statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ and the total uncertainty $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ including the contributions of the radiative corrections, see (5.2.1). An additional luminosity error of $\delta\left(\mathcal{L}_{data}\right) = 2\%$ needs to be added to obtain the full uncertainty.*
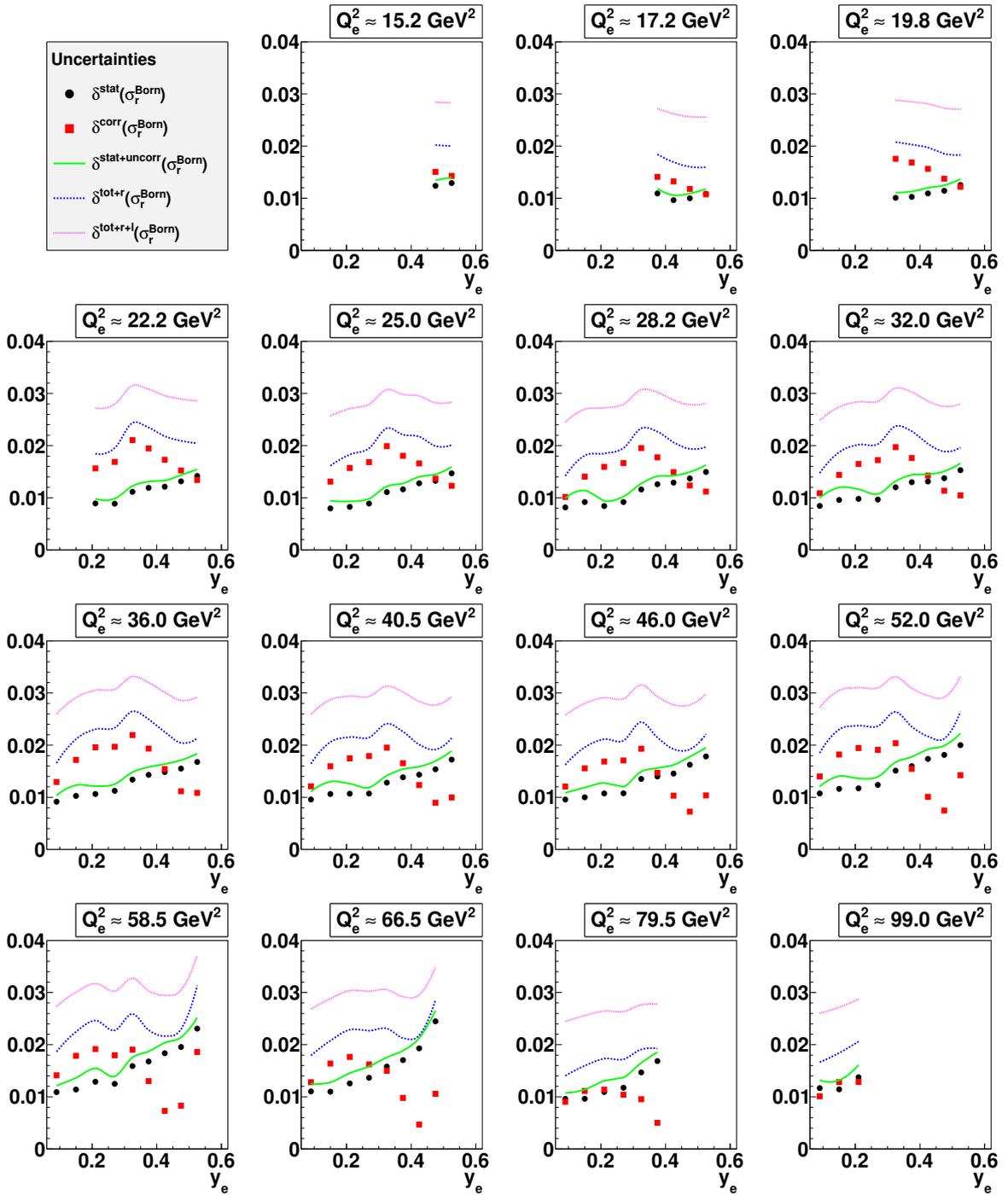
Figure 12.3.6: *Different contributions to the total relative uncertainty. The markers shows the statistical uncertainty $\delta^{stat}\left(\sigma_r^{Born}\right)$ and the cumulated error $\delta^{corr}\left(\sigma_r^{Born}\right)$ of all correlated systematic error sources. The lines depict cumulated errors: $\delta^{stat+uncorr}\left(\sigma_r^{Born}\right)$ contains the statistical error and the total uncertainty of all uncorrelated error sources. The total error $\delta^{tot}\left(\sigma_r^{Born}\right)$ also includes the correlated error sources, while for $\delta^{tot+r}\left(\sigma_r^{Born}\right)$ also the uncertainty from the radiative correction is added. $\delta^{tot+r+l}\left(\sigma_r^{Born}\right)$ contains all error sources including the luminosity error.*
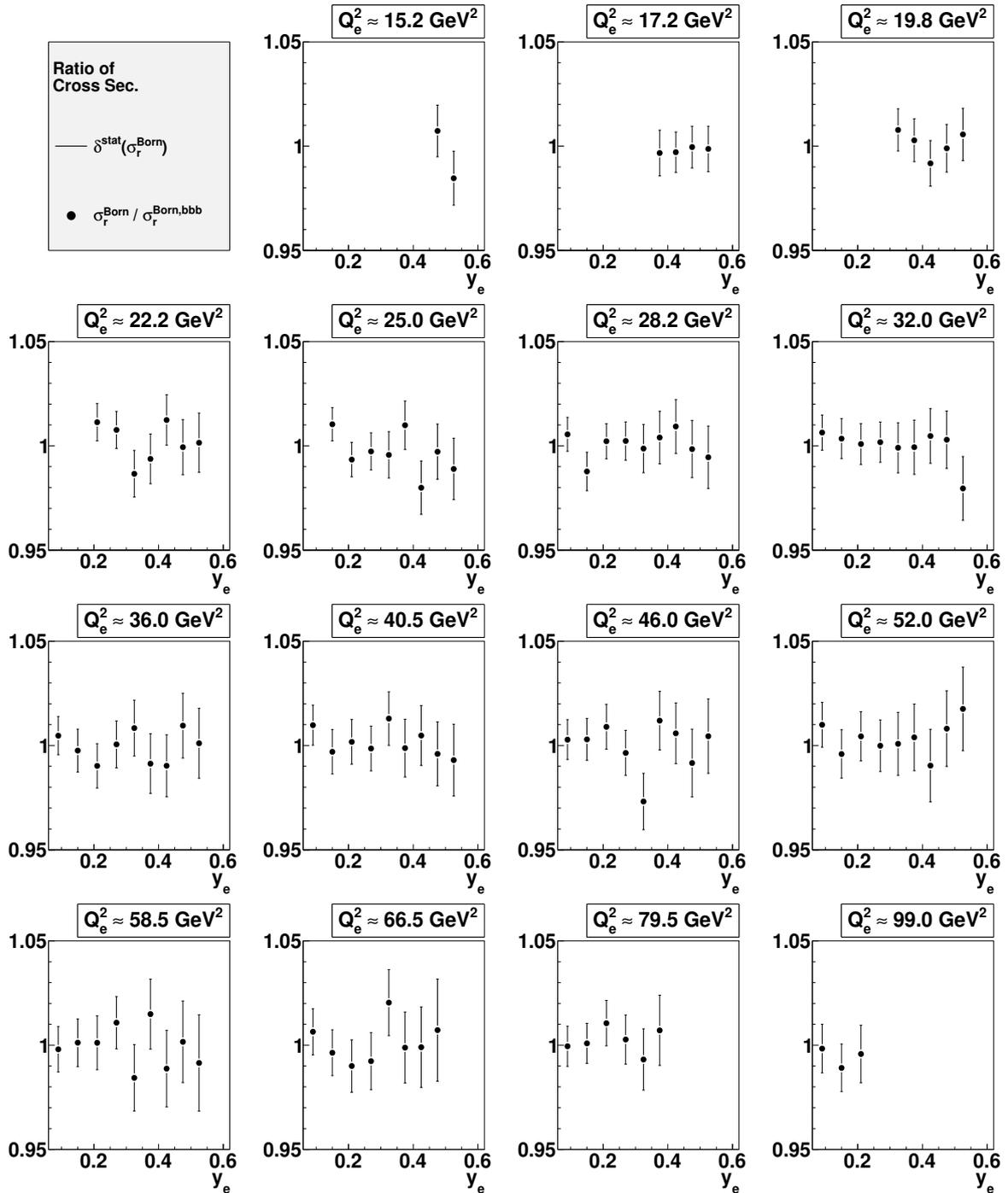
Figure 12.3.7: *Ratio plot of σ_r^{Born} to the unfolding result and of the bin-by-bin method $\sigma_r^{Born,bbb}$. The relative statistical uncertainties $\delta^{stat}\left(\sigma_r^{Born}\right)$ are drawn.*
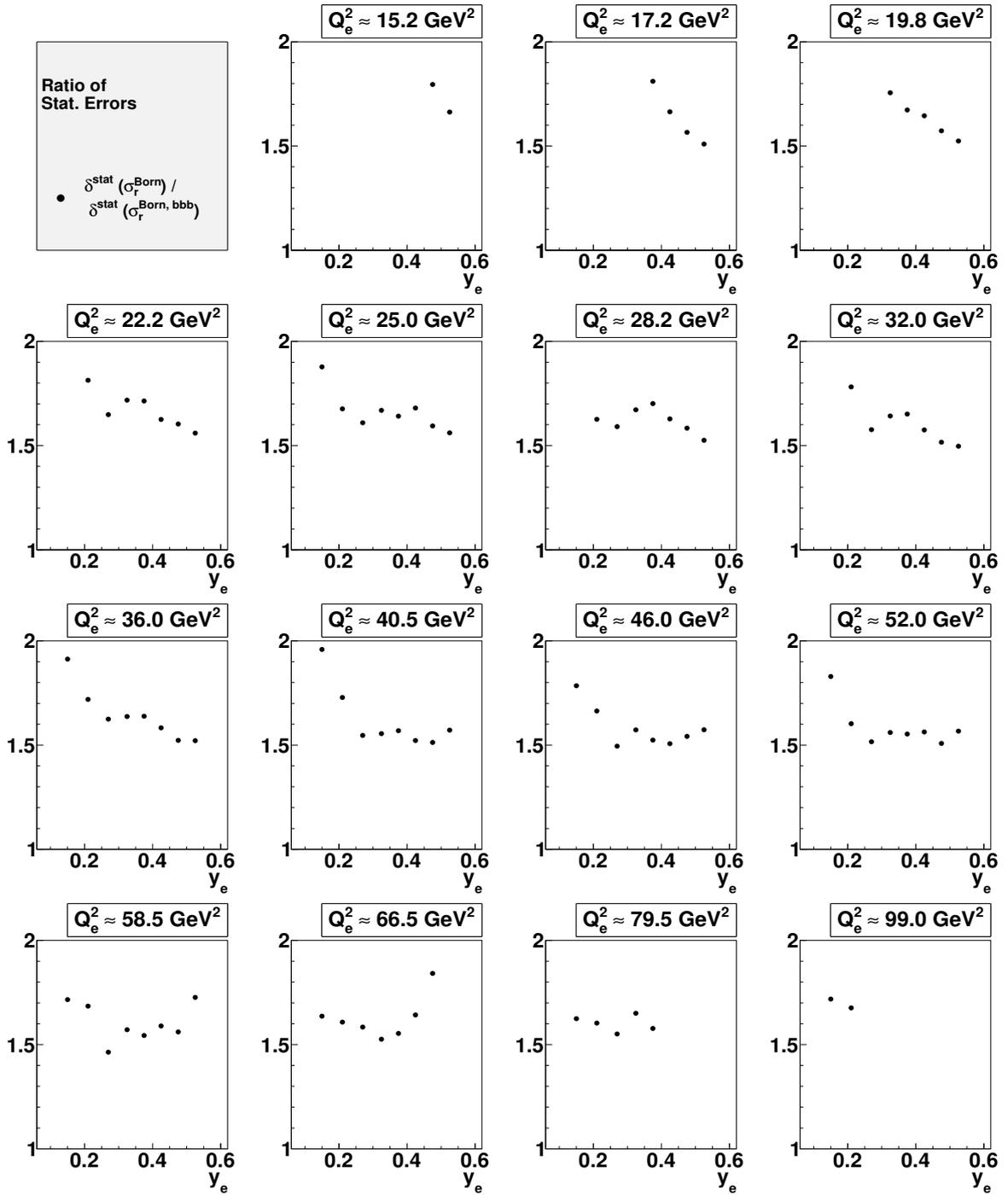
Figure 12.3.8: *Ratio of statistical uncertainties from unfolding ($\delta^{stat}(\sigma_r^{Born})$) and from the bin-by-bin method $\delta^{stat}(\sigma_r^{Born,bbb})$.*
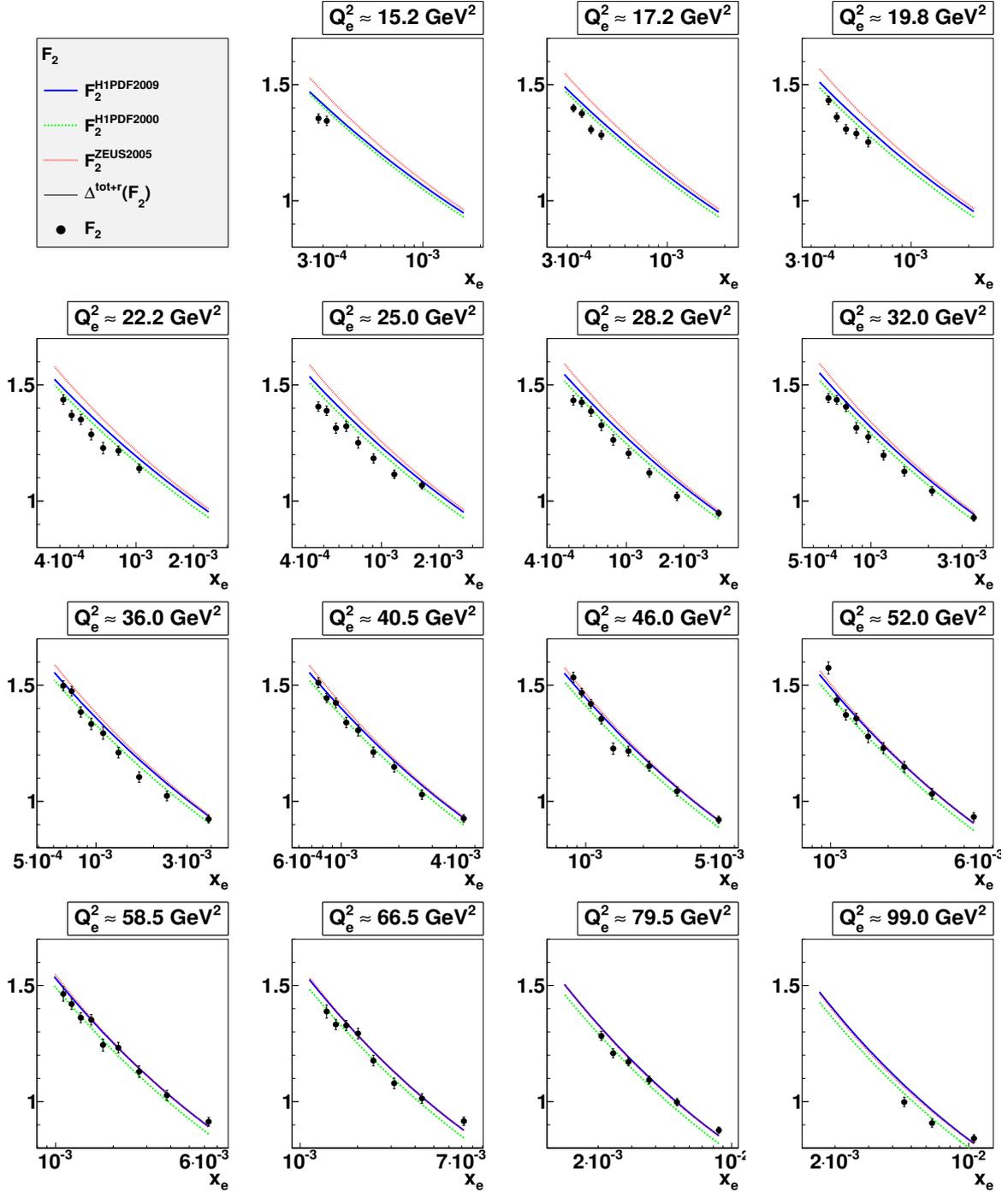
Figure 12.3.9: *The structure function F_2. The error bars $\Delta^{tot+r}\left(F_2\left(x_e, Q_e^2\right)\right)$ contain the statistical uncertainty and the total uncertainties including radiative corrections. An additional luminosity uncertainty of 2% has to be applied. For comparison, the expected shape according to the pdf sets* $\mathtt{H1PDF2009}$ *($F_2^{H1PDF09}$)* [Kre, A+09], $\mathtt{H1PDF2000}$ *($F_2^{H1PDF00}$)* [A+03] *and* $\mathtt{ZEUS2005}$ *($F_2^{ZEUS2005}$) is plotted* [C+03].

Chapter 13

Conclusion

In this analysis the double differential inclusive neutral current $ep \to eX$ cross section has been measured for 106 data points in the kinematic range of $0.06 < y < 0.6$ for the inelasticity and $14 \text{ GeV}^2 < Q_e^2 < 110 \text{ GeV}^2$ for the squared momentum exchange.

In previous measurements, detector effects such as migration and inefficiency have been treated by a bin-by-bin efficiency correction. In this analysis, two dimensional unfolding in y and Q^2 has been applied for the first time. Different regularization methods have been systematically tested and compared. In particular, the algebraic method, the global correlation method and the L curve method have been studied. All three methods give similar results, which are consistent with the result from the standard bin-by-bin method. However, the statistical uncertainties from unfolding are larger than those from the standard bin-by-bin method for all tested regularization prescriptions. Differences on the systematic uncertainties are pronounced in regions of large variations of the data distribution.

The global correlation method gives the smoothest solution which minimizes the correlation between the data points. In this case, a difference of the order of $20 - 30\%$ is found for the statistical uncertainties in comparison to the bin-by-bin correction method. This reflects the effect of migration on the statistical error, a contribution that is not accounted for by the traditional (bin-by-bin) error treatment. The bin-by-bin method indeed underestimates the statistical uncertainty. However, for these smooth distributions the transition from the bin-by-bin method to the unfolding method does not affect the mean values of the cross section measurement.

For the algebraic method, the statistical uncertainty is of the order of $1 - 2\%$ and exhibits a rising trend with y and Q^2. The correlated systematic uncertainty of this analysis takes on values of $1 - 2\%$ as well. At values of small inelasticity y it is dominated by the scale uncertainty on the electron energy E_e, where it takes on values around $\delta^{E_e}(\sigma_r) \simeq 2\%$. For small values of the squared momentum exchange Q^2, the uncertainty of the polar angle measurement dominates, giving values of $\delta^{E_\theta}(\sigma_r) \simeq 2\%$. The total uncertainty including all statistical and systematic error sources takes on values between $2 - 3\%$ throughout the kinematic range of this analysis. In addition, a luminosity uncertainty of $\delta(\mathcal{L}_{data}) \simeq 2\%$ must be considered.

The results are compatible with the pdf fit `H1PDF2000` [A+03], although the ratio exhibits a slight rising trend with both y and Q^2. Similar statements can be made for the comparison with `H1PDF2009` [Kre, A+09], which however exhibits an additional normalization difference of $2 - 3\%$. This difference is not an artefact of the unfolding

method.

This work improves the understanding of both statistical and systematic uncertainties of the inclusive cross section measurement in the domain of the SpaCal calorimeter. Moreover, the bin-to-bin correlations are assessed for the first time. The results and correlations obtained here lend themselves for a systematic analysis of the structure function and its overall normalization to improve the extraction of the PDF. That additional step, however, was beyond the scope of this analysis.

Appendix A

Cross Sections, Uncertainties and Correlations

The following pages contain the measured reduced cross sections in tabular form. Only bins with sufficient efficiency ($\epsilon > 70\%$) and no overflow bins are shown. The level of regularization corresponds to the algebraical method, i. e. $\log \tau_A = -4.29$.

For each bin $\mathcal{B}_i$, both the radiative and the Born level cross section are given (σ_r^{rad} and σ_r^{Born}). Moreover, the functional value of the structure fuction F_2 is given. Each of these values is related to a bin center (Q_e^2, y_e, x_e) which is given in the last three columns of the table. The values of Q_e^2 and y_e have been calculated with (5.3.5) and (5.3.6), then x_e is obtained from (2.1.7).

The relative uncertainties are given for the following combination of error sources:

- Relative statistical uncertainty $\delta^{stat}\left(\sigma_{red}^{Born}\right)$, according to (4.8.5).

- Relative systematic uncertainty $\delta^{uncorr}\left(\sigma_{red}^{Born}\right)$ from uncorrelated error sources, see (4.8.12).

- Relative systematic uncertainty $\delta^{corr}\left(\sigma_{red}^{Born}\right)$ from all correlated error sources, added up in quadrature. See (4.8.19) and (4.8.22) for the systematic uncertainty of a single error source λ.

- Total relative systematic uncertainty $\delta^{tot}\left(\sigma_{red}^{Born}\right)$ comprising of all error sources except radiative corrections and luminosity.

- Total relative systematic uncertainty $\delta^{tot+r}\left(\sigma_{red}^{Born}\right)$ comprising of all error sources including radiative corrections but except luminosity.

- Total relative systematic uncertainty $\delta^{tot+r+l}\left(\sigma_{red}^{Born}\right)$ comprising of all error sources including radiative corrections and luminosity.

Moreover, the global correlation ρ_i is given for each bin $\mathcal{B}_i$.

Due to the amount of data, the bin-to-bin correlations ρ_{ij} have been suppressed in this table. They can be obtained in electronic form from the author.

Bin $\mathcal{B}_i$	σ_r^{rad}	σ_r^{Born}	F_2	δ^{stat} $\left(\sigma_r^{Born}\right)$ [%]	δ^{uncorr} $\left(\sigma_r^{Born}\right)$ [%]	δ^{corr} $\left(\sigma_r^{Born}\right)$ [%]	δ^{tot} $\left(\sigma_r^{Born}\right)$ [%]	δ^{tot+r} $\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r+l}$ $\left(\sigma_r^{Born}\right)$ [%]	ρ	Q_e^2 [GeV^2]	y_e	x_e
28	1.062	0.947	0.948	0.81	0.28	1.02	1.42	1.42	2.45	0.279	27.82	0.090	0.0030
29	1.041	0.928	0.929	0.84	0.18	1.09	1.48	1.48	2.48	0.183	31.52	0.090	0.0030
30	1.030	0.923	0.924	0.92	0.24	1.29	1.66	1.66	2.60	0.242	35.48	0.090	0.0040
31	1.029	0.925	0.926	0.96	0.20	1.21	1.64	1.64	2.59	0.198	39.91	0.090	0.0040
32	1.022	0.920	0.921	0.96	0.17	1.21	1.62	1.62	2.58	0.175	45.33	0.090	0.0050
33	1.031	0.933	0.933	1.07	0.15	1.40	1.85	1.85	2.72	0.152	51.27	0.090	0.0060
34	1.007	0.913	0.914	1.09	0.19	1.41	1.87	1.87	2.73	0.192	57.69	0.090	0.0060
35	1.009	0.915	0.916	1.10	0.19	1.28	1.79	1.79	2.69	0.188	65.56	0.090	0.0070
36	0.962	0.876	0.877	0.96	0.22	0.90	1.40	1.40	2.44	0.223	78.11	0.090	0.0090
37	0.918	0.841	0.842	1.17	0.00	1.01	1.66	1.66	2.60	0.000	97.25	0.090	0.0110
47	1.271	1.065	1.068	0.80	0.44	1.31	1.62	1.62	2.57	0.439	24.83	0.150	0.0020
48	1.211	1.018	1.021	0.92	0.44	1.40	1.81	1.81	2.70	0.445	28.05	0.150	0.0020
49	1.232	1.041	1.044	0.96	0.37	1.44	1.87	1.87	2.74	0.371	31.78	0.150	0.0020
50	1.209	1.021	1.024	1.03	0.42	1.72	2.11	2.11	2.91	0.417	35.77	0.150	0.0020
51	1.209	1.027	1.030	1.06	0.35	1.60	2.06	2.06	2.87	0.349	40.23	0.150	0.0030
52	1.216	1.041	1.044	1.00	0.38	1.56	1.96	1.96	2.80	0.379	45.68	0.150	0.0030
53	1.207	1.030	1.032	1.16	0.34	1.82	2.30	2.30	3.05	0.339	51.67	0.150	0.0030
54	1.201	1.025	1.027	1.14	0.29	1.79	2.25	2.25	3.01	0.287	58.12	0.150	0.0040
55	1.175	1.011	1.014	1.10	0.34	1.64	2.08	2.08	2.88	0.341	66.04	0.150	0.0040
56	1.149	0.995	0.998	0.96	0.31	1.12	1.60	1.60	2.56	0.312	78.68	0.150	0.0050
57	1.049	0.906	0.908	1.14	0.00	1.28	1.84	1.84	2.72	0.000	97.95	0.150	0.0060
66	1.407	1.135	1.141	0.89	0.26	1.56	1.85	1.85	2.72	0.259	22.15	0.210	0.0010
67	1.371	1.109	1.115	0.83	0.23	1.57	1.83	1.83	2.71	0.228	24.89	0.210	0.0010
68	1.373	1.115	1.121	0.84	0.35	1.59	1.85	1.85	2.72	0.354	28.12	0.210	0.0010
69	1.374	1.122	1.128	0.98	0.27	1.65	2.02	2.02	2.84	0.267	31.85	0.210	0.0010
70	1.341	1.099	1.105	1.06	0.33	1.96	2.31	2.31	3.05	0.329	35.85	0.210	0.0020

Bin $\mathcal{B}_i$	σ_r^{rad}	σ_r^{Born}	F_2	δ^{stat} $\left(\sigma_r^{Born}\right)$ [%]	δ^{uncorr} $\left(\sigma_r^{Born}\right)$ [%]	δ^{corr} $\left(\sigma_r^{Born}\right)$ [%]	δ^{tot} $\left(\sigma_r^{Born}\right)$ [%]	δ^{tot+r} $\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r+l}$ $\left(\sigma_r^{Born}\right)$ [%]	ρ	Q_e^2 [GeV^2]	y_e	x_e
71	1.384	1.143	1.149	1.07	0.32	1.75	2.15	2.15	2.94	0.320	40.31	0.210	0.0020
72	1.392	1.146	1.152	1.08	0.26	1.69	2.11	2.11	2.91	0.261	45.78	0.210	0.0020
73	1.384	1.142	1.148	1.17	0.34	1.94	2.37	2.37	3.10	0.336	51.78	0.210	0.0020
74	1.352	1.124	1.129	1.29	0.26	1.92	2.46	2.46	3.17	0.264	58.25	0.210	0.0030
75	1.296	1.073	1.079	1.26	0.34	1.76	2.29	2.29	3.04	0.343	66.18	0.210	0.0030
76	1.313	1.088	1.093	1.09	0.27	1.14	1.73	1.73	2.65	0.267	78.84	0.210	0.0040
77	1.195	0.993	0.998	1.38	0.00	1.28	2.06	2.06	2.87	0.000	98.13	0.210	0.0050
86	1.525	1.206	1.217	0.89	0.24	1.69	1.96	1.96	2.80	0.236	22.17	0.270	0.0010
87	1.488	1.173	1.184	0.89	0.24	1.68	1.95	1.95	2.79	0.242	24.91	0.270	0.0010
88	1.512	1.194	1.205	0.92	0.23	1.66	1.95	1.95	2.80	0.230	28.15	0.270	0.0010
89	1.496	1.186	1.197	0.97	0.24	1.72	2.03	2.03	2.85	0.241	31.88	0.270	0.0010
90	1.503	1.199	1.210	1.12	0.22	1.97	2.33	2.33	3.07	0.215	35.88	0.270	0.0010
91	1.501	1.201	1.212	1.07	0.21	1.79	2.15	2.15	2.94	0.207	40.35	0.270	0.0010
92	1.507	1.206	1.217	1.08	0.21	1.71	2.09	2.09	2.89	0.213	45.82	0.270	0.0020
93	1.514	1.218	1.229	1.24	0.19	1.91	2.37	2.37	3.10	0.191	51.83	0.270	0.0020
94	1.519	1.222	1.232	1.25	0.27	1.80	2.27	2.27	3.03	0.272	58.30	0.270	0.0020
95	1.446	1.166	1.177	1.37	0.29	1.62	2.27	2.27	3.02	0.290	66.24	0.270	0.0020
96	1.424	1.161	1.171	1.17	0.00	1.04	1.72	1.72	2.64	0.000	78.91	0.270	0.0030
105	1.604	1.236	1.253	1.01	0.24	1.76	2.08	2.08	2.88	0.243	19.69	0.325	0.0010
106	1.564	1.211	1.229	1.11	0.24	2.11	2.44	2.44	3.15	0.242	22.19	0.325	0.0010
107	1.591	1.234	1.252	1.11	0.24	1.99	2.33	2.33	3.07	0.243	24.93	0.325	0.0010
108	1.609	1.245	1.263	1.16	0.23	1.95	2.33	2.33	3.07	0.234	28.17	0.325	0.0010
109	1.610	1.258	1.276	1.20	0.23	1.97	2.37	2.37	3.10	0.230	31.91	0.325	0.0010
110	1.634	1.276	1.294	1.34	0.21	2.19	2.65	2.65	3.32	0.212	35.91	0.325	0.0010
111	1.638	1.288	1.306	1.28	0.20	1.95	2.41	2.41	3.13	0.203	40.39	0.325	0.0010
112	1.548	1.210	1.228	1.35	0.20	1.93	2.44	2.44	3.16	0.198	45.86	0.325	0.0010

Bin $\mathcal{B}_i$	σ_r^{rad}	σ_r^{Born}	F_2	$\delta^{stat}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{uncorr}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{corr}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r+l}\left(\sigma_r^{Born}\right)$ [%]	ρ	Q_e^2 [GeV^2]	y_e	x_e
113	1.605	1.262	1.279	1.51	0.19	2.04	2.64	2.64	3.31	0.195	51.87	0.325	0.0020
114	1.560	1.227	1.244	1.59	0.21	1.90	2.59	2.59	3.27	0.212	58.35	0.325	0.0020
115	1.622	1.276	1.293	1.58	0.25	1.50	2.31	2.31	3.06	0.245	66.29	0.325	0.0020
116	1.504	1.192	1.209	1.47	0.00	0.95	1.91	1.91	2.76	0.000	78.97	0.325	0.0020
124	1.660	1.259	1.283	1.09	0.24	1.41	1.84	1.84	2.72	0.242	17.19	0.375	0.0000
125	1.670	1.264	1.289	1.03	0.25	1.68	2.03	2.03	2.85	0.253	19.70	0.375	0.0010
126	1.668	1.262	1.287	1.19	0.24	1.95	2.35	2.35	3.08	0.240	22.20	0.375	0.0010
127	1.706	1.296	1.322	1.16	0.28	1.80	2.21	2.21	2.98	0.278	24.94	0.375	0.0010
128	1.693	1.300	1.326	1.26	0.25	1.77	2.27	2.27	3.02	0.251	28.18	0.375	0.0010
129	1.689	1.289	1.316	1.30	0.22	1.76	2.28	2.28	3.03	0.223	31.92	0.375	0.0010
130	1.695	1.307	1.333	1.43	0.22	1.93	2.50	2.50	3.20	0.224	35.92	0.375	0.0010
131	1.717	1.314	1.340	1.38	0.22	1.65	2.26	2.26	3.02	0.217	40.40	0.375	0.0010
132	1.735	1.329	1.355	1.40	0.21	1.47	2.14	2.14	2.93	0.208	45.87	0.375	0.0010
133	1.716	1.331	1.357	1.60	0.22	1.54	2.35	2.35	3.09	0.216	51.88	0.375	0.0010
134	1.714	1.327	1.353	1.68	0.21	1.30	2.28	2.28	3.03	0.211	58.35	0.375	0.0020
135	1.664	1.303	1.328	1.71	0.16	0.98	2.13	2.13	2.92	0.155	66.30	0.375	0.0020
136	1.624	1.259	1.283	1.68	0.00	0.50	1.92	1.92	2.77	0.000	78.98	0.375	0.0020
144	1.714	1.273	1.307	0.96	0.23	1.32	1.69	1.69	2.62	0.231	17.19	0.425	0.0000
145	1.715	1.274	1.308	1.09	0.23	1.57	1.97	1.97	2.81	0.229	19.70	0.425	0.0000
146	1.764	1.316	1.351	1.21	0.24	1.73	2.18	2.18	2.96	0.242	22.20	0.425	0.0010
147	1.705	1.279	1.314	1.28	0.24	1.66	2.17	2.17	2.95	0.241	24.94	0.425	0.0010
148	1.804	1.350	1.387	1.29	0.23	1.49	2.06	2.06	2.87	0.228	28.18	0.425	0.0010
149	1.814	1.370	1.406	1.31	0.21	1.43	2.03	2.03	2.85	0.207	31.92	0.425	0.0010
150	1.778	1.349	1.385	1.49	0.20	1.54	2.25	2.25	3.01	0.204	35.92	0.425	0.0010
151	1.828	1.388	1.424	1.43	0.20	1.24	2.01	2.01	2.84	0.203	40.40	0.425	0.0010
152	1.833	1.384	1.420	1.46	0.20	1.03	1.92	1.92	2.78	0.197	45.88	0.425	0.0010

Bin $\mathcal{B}_i$	σ_r^{rad}	σ_r^{Born}	F_2	δ^{stat} $\left(\sigma_r^{Born}\right)$ [%]	δ^{uncorr} $\left(\sigma_r^{Born}\right)$ [%]	δ^{corr} $\left(\sigma_r^{Born}\right)$ [%]	δ^{tot} $\left(\sigma_r^{Born}\right)$ [%]	δ^{tot+r} $\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r+l}$ $\left(\sigma_r^{Born}\right)$ [%]	ρ	Q_e^2 [GeV^2]	y_e	x_e
153	1.759	1.336	1.372	1.74	0.21	1.01	2.17	2.17	2.95	0.213	51.88	0.425	0.0010
154	1.749	1.326	1.361	1.84	0.20	0.73	2.16	2.16	2.95	0.196	58.36	0.425	0.0010
155	1.712	1.298	1.332	1.93	0.00	0.47	2.18	2.18	2.96	0.000	66.31	0.425	0.0020
163	1.790	1.301	1.344	1.24	0.21	1.50	2.02	2.02	2.84	0.214	15.12	0.475	0.0000
164	1.808	1.331	1.376	1.00	0.20	1.18	1.61	1.61	2.56	0.204	17.20	0.475	0.0000
165	1.804	1.314	1.360	1.14	0.21	1.37	1.86	1.86	2.73	0.212	19.70	0.475	0.0000
166	1.806	1.322	1.369	1.32	0.22	1.52	2.09	2.09	2.89	0.220	22.20	0.475	0.0000
167	1.820	1.341	1.389	1.32	0.21	1.37	1.99	1.99	2.82	0.210	24.94	0.475	0.0010
168	1.860	1.378	1.426	1.37	0.19	1.24	1.94	1.94	2.79	0.192	28.18	0.475	0.0010
169	1.872	1.387	1.436	1.38	0.19	1.14	1.89	1.89	2.75	0.190	31.92	0.475	0.0010
170	1.931	1.426	1.475	1.55	0.18	1.12	2.04	2.04	2.86	0.177	35.93	0.475	0.0010
171	1.894	1.396	1.445	1.54	0.16	0.89	1.92	1.92	2.77	0.157	40.40	0.475	0.0010
172	1.902	1.419	1.468	1.62	0.17	0.73	1.92	1.92	2.77	0.165	45.88	0.475	0.0010
173	1.897	1.388	1.436	1.81	0.20	0.75	2.13	2.13	2.92	0.196	51.89	0.475	0.0010
174	1.837	1.373	1.420	1.96	0.12	0.83	2.30	2.30	3.05	0.125	58.37	0.475	0.0010
175	1.792	1.341	1.388	2.45	0.00	1.06	2.86	2.86	3.49	0.000	66.32	0.475	0.0010
183	1.807	1.298	1.355	1.29	0.15	1.43	2.00	2.00	2.83	0.147	15.12	0.525	0.0000
184	1.878	1.340	1.399	1.09	0.15	1.07	1.60	1.60	2.56	0.147	17.20	0.525	0.0000
185	1.901	1.372	1.433	1.25	0.15	1.22	1.83	1.83	2.71	0.150	19.70	0.525	0.0000
186	1.898	1.375	1.437	1.42	0.15	1.34	2.05	2.05	2.86	0.147	22.21	0.525	0.0000
187	1.853	1.344	1.407	1.47	0.14	1.23	2.01	2.01	2.84	0.142	24.94	0.525	0.0000
188	1.880	1.371	1.434	1.50	0.13	1.12	1.97	1.97	2.81	0.126	28.18	0.525	0.0010
189	1.898	1.380	1.444	1.53	0.12	1.05	1.96	1.96	2.80	0.119	31.92	0.525	0.0010
190	1.984	1.434	1.498	1.68	0.11	1.09	2.13	2.13	2.92	0.110	35.93	0.525	0.0010
191	1.973	1.447	1.511	1.72	0.10	1.00	2.14	2.14	2.93	0.100	40.41	0.525	0.0010
192	2.010	1.470	1.534	1.78	0.16	1.04	2.21	2.21	2.98	0.163	45.88	0.525	0.0010

Bin $\mathcal{B}_i$	σ_r^{rad}	σ_r^{Born}	F_2	$\delta^{stat}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{uncorr}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{corr}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r}\left(\sigma_r^{Born}\right)$ [%]	$\delta^{tot+r+l}\left(\sigma_r^{Born}\right)$ [%]	ρ	Q_e^2 [GeV^2]	y_e	x_e
193	2.021	1.511	1.574	2.00	0.11	1.42	2.64	2.64	3.31	0.112	51.89	0.525	0.0010
194	1.912	1.402	1.464	2.31	0.00	1.86	3.13	3.13	3.72	0.000	58.37	0.525	0.0010

Bibliography

[A+92]   T. C. Awes et al. A Simple method of shower localization and identification in laterally segmented calorimeters. *Nucl. Instrum. Meth.*, A311:130–138, 1992.

[A+93a] B. Andrieu et al. Results from pion calibration runs for the H1 liquid argon calorimeter and comparisons with simulations. *Nucl. Instrum. Meth.*, A336:499–509, 1993.

[A+93b] B. Andrieu et al. The H1 liquid argon calorimeter system. *Nucl. Instrum. Meth.*, A336:460–498, 1993.

[A+96]   R. D. Appuhn et al. Hadronic response and e/pi separation with the H1 lead/fibre calorimeter. *Nucl. Instrum. Meth.*, A382:395–412, 1996.

[A+97a] I. Abt et al. The H1 detector at HERA. *Nucl. Instrum. Meth.*, A386:310–347, 1997.

[A+97b] M. Arneodo et al. Measurement of the proton and deuteron structure functions, F2(p) and F2(d), and of the ratio sigma(L)/sigma(T). *Nucl. Phys.*, B483:3–43, 1997.

[A+03]   C. Adloff et al. Measurement and QCD analysis of neutral and charged current cross sections at HERA. *Eur. Phys. J.*, C30:1–32, 2003.

[A+08]   F. D. Aaron et al. Measurement of the Proton Structure Function F_L at Low x. *Phys. Lett.*, B665:139–146, 2008.

[A+09]   F. D. Aaron et al. A Precision Measurement of the Inclusive ep Scattering Cross Section at HERA. *Eur. Phys. J.*, C64:561–587, 2009.

[A+10]   F. D. Aaron et al. Combined Measurement and QCD Analysis of the Inclusive ep Scattering Cross Sections at HERA. *JHEP*, 01:109, 2010.

[ABB+96] A. Arbuzov, Dmitri Yu. Bardin, J. Blumlein, L. Kalinovskaya, and T. Riemann. HECTOR 1.00 - A program for the calculation of QED, QCD and electroweak corrections to ep and lN deep inelastic neutral and charged current scattering. *Comput. Phys. Commun.*, 94:128–184, 1996.

[AGIS83] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton Fragmentation and String Dynamics. *Phys. Rept.*, 97:31–145, 1983.

[AGLP89] Bo Andersson, Gosta Gustafson, Leif Lonnblad, and Ulf Pettersson. Coherence Effects in Deep Inelastic Scattering. *Z. Phys.*, C43:625, 1989.

[AGS83] Bo Andersson, G. Gustafson, and B. Soderberg. A General Model for Jet Fragmentation. *Z. Phys.*, C20:317, 1983.

[AP77] Guido Altarelli and G. Parisi. Asymptotic Freedom in Parton Language. *Nucl. Phys.*, B126:298, 1977.

[B⁺] Raymond Brock et al. Handbook of perturbative QCD; Version 1.1: September 1994. FERMILAB-PUB-94-316.

[B⁺89] A. C. Benvenuti et al. A High Statistics Measurement of the Proton Structure Functions F(2) (x, Q**2) and R from Deep Inelastic Muon Scattering at High Q**2. *Phys. Lett.*, B223:485, 1989.

[B⁺95] Raymond Brock et al. Handbook of perturbative QCD: Version 1.0. *Rev. Mod. Phys.*, 67:157–248, 1995.

[B⁺08] J. Becker et al. A vertex trigger based on cylindrical multiwire proportional chambers. *Nucl. Instrum. Meth.*, A586:190–203, 2008.

[BB95] Ursula Bassler and Gregorio Bernardi. On the kinematic reconstruction of deep inelastic scattering at HERA: The Sigma method. *Nucl. Instrum. Meth.*, A361:197–208, 1995.

[BBM⁺] R. Brun, F. Bruyant, M. Maire, A. C. McPherson, and P. Zanarini. GEANT3. CERN-DD-EE-84-1.

[Biz97] J. C. et al. Bizot. Strategy studies for the h1 topological l2 trigger (l2tt). 1997.

[Bjo69] J. D. Bjorken. Asymptotic Sum Rules at Infinite Momentum. *Phys. Rev.*, 179:1547–1553, 1969.

[Bloa] V. Blobel. Unfolding Methods in High-Energy Physics Experiments. Lectures given at 1984 CERN School of Computing, Aiguablava, Spain, Sep 9-22, 1984.

[Blob] Volker Blobel. Private conversation.

[Bloc] Volker Blobel. Some comments on chi**2 minimisation applications. Prepared for PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology, Menlo Park, California, 8-11 Sep 2003.

[Blo01] Volker Blobel. Statistical formulas. www.desy.de/ blobel/statform.ps, 10 2001.

[Blo02] Volker Blobel. An unfolding method for high energy physics experiments. *hep-ex/0208022*, 2002.

[Blo06] V. Blobel. A new fast track-fit algorithm based on broken lines. *Nucl. Instrum. Meth.*, A566:14–17, 2006.

[Blu91] Johannes Bluemlein. Leading log radiative corrections to e p scattering including jet measurement. *Phys. Lett.*, B271:267–273, 1991.

[C⁺03] S. Chekanov et al. A ZEUS next-to-leading-order QCD analysis of data on deep inelastic scattering. *Phys. Rev.*, D67:012007, 2003.

[CG69] Curtis G. Callan, Jr. and David J. Gross. High-energy electroproduction and the constitution of the electric current. *Phys. Rev. Lett.*, 22:156–159, 1969.

[Col93] ZEUS Collaboration. The ZEUS detector: Status report 1993. 1993. ZEUS-STATUS-REPT-1993.

[Col09] H1 Collaboration. Measurement of the Inclusive ep Scattering Cross Section at Low Q^2 and x at HERA. *Eur. Phys. J.*, C63:625–678, 2009.

[Cow] G. Cowan. A survey of unfolding methods for particle physics. Prepared for Conference on Advanced Statistical Techniques in Particle Physics, Durham, England, 18-22 Mar 2002.

[CS03] Amanda Cooper-Sarkar. *Deep Inelastic Scattering.* Oxford University Press, 2003.

[CS10] Amanda Cooper-Sarkar. Proton Structure from HERA to LHC. *arxiv.org, eprint 1012.1438*, 2010.

[CSS88] John C. Collins, Davison E. Soper, and George F. Sterman. Factorization of Hard Processes in QCD. *Adv. Ser. Direct. High Energy Phys.*, 5:1–91, 1988.

[D'A95] G. D'Agostini. A Multidimensional unfolding method based on Bayes' theorem. *Nucl. Instrum. Meth.*, A362:487–498, 1995.

[Dok77] Yuri L. Dokshitzer. Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics. *Sov. Phys. JETP*, 46:641–653, 1977.

[ES] R. Keith Ellis and W. James Stirling. QCD and collider physics. Based on lectures given at CERN School of Physics, Majorca, Spain, Sep 16-29, 1990 and CERN-JINR School of Physics, Egmond, Netherlands, Jun 25 - Jul 8, 1989.

[F⁺74] D. J. Fox et al. Test of Scale Invariance in High-Energy Muon Scattering. *Phys. Rev. Lett.*, 33:1504, 1974.

[FC98] Gary J. Feldman and Robert D. Cousins. A Unified Approach to the Classical Statistical Analysis of Small Signals. *Phys. Rev.*, D57:3873–3889, 1998.

[GL72a] V. N. Gribov and L. N. Lipatov. Deep inelastic e p scattering in perturbation theory. *Sov. J. Nucl. Phys.*, 15:438–450, 1972.

[GL72b] V. N. Gribov and L. N. Lipatov. e+ e- pair annihilation and deep inelastic e p scattering in perturbation theory. *Sov. J. Nucl. Phys.*, 15:675–684, 1972.

[Gla] Alexandre Alimovich Glazov. Measurement of the proton structure functions F2(x,Q**2) and F(L)(x,Q**2) with the H1 detector at HERA. DESY-THESIS-1998-005.

[Gla06] A. Glazov. Low q2 hera-2 issues. ELAN Talk May 16, 2006, https://www-h1.desy.de/icgi-h1wiki/moin.cgi/ElanInclusive/Meeting2006-05-16, 2006.

[Gla08] A. Glazov. Recent measurements of the proton structure functions at HERA. *Acta Phys. Polon. Supp.*, 1:371–378, 2008.

[GP88] Gosta Gustafson and Ulf Pettersson. Dipole Formulation of QCD Cascades. *Nucl. Phys.*, B306:746, 1988.

[GRP90] Guenter Grindhammer, M. Rudowicz, and S. Peters. THE FAST SIMULATION OF ELECTROMAGNETIC AND HADRONIC SHOWERS. *Nucl. Instrum. Meth.*, A290:469, 1990.

[Han00] P. C. Hansen. The l-curve and its use in the numerical treatment of inverse problems. In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

[HK96] Andreas Hocker and Vakhtang Kartvelishvili. SVD Approach to Data Unfolding. *Nucl. Instrum. Meth.*, A372:469–481, 1996.

[IER97] G. Ingelman, A. Edin, and J. Rathsman. LEPTO 6.5 - A Monte Carlo Generator for Deep Inelastic Lepton-Nucleon Scattering. *Comput. Phys. Commun.*, 101:108–134, 1997.

[Jam] Frederick James. Statistical methods in experimental physics. ISBN-9789812567956.

[Jun07] A. W. et al. Jung. Proc. of the 2007 ieee npss realtime conference, isbn 1-4244-0867-9, (oct. 2007), p. ieee catalog number: 07ex1643c. 2007.

[K+97] J. K. Kohne et al. Realization of a second level neural network trigger for the H1 experiment at HERA. *Nucl. Instrum. Meth.*, A389:128–133, 1997.

[Kre] Jan Kretzschmar. A precision measurement of the proton structure function F2 with the H1 experiment,.

[KSM92] A. Kwiatkowski, H. Spiesberger, and H. J. Mohring. HERACLES: An Event Generator for ep Interactions at HERA Energies Including Radiative Processes: Version 1.0. *Comp. Phys. Commun.*, 69:155–172, 1992.

[L+00] H. L. Lai et al. Global QCD analysis of parton structure of the nucleon: CTEQ5 parton distributions. *Eur. Phys. J.*, C12:375–392, 2000.

[Lon92] Leif Lonnblad. ARIADNE version 4: A Program for simulation of QCD cascades implementing the color dipole model. *Comput. Phys. Commun.*, 71:15–31, 1992.

[M$^+$72] Guthrie Miller et al. Inelastic electron-Proton Scattering at Large Momentum Transfers. *Phys. Rev.*, D5:528, 1972.

[MEP95] Daniel V. Schröder Michael E. Peskin. *An Introduction to Quantum Field Theory.* Perseus Books, 1995.

[Mey] Andreas B. Meyer. Heavy quark production at HERA. Habilitationsschrift.

[MVV04] S. Moch, J. A. M. Vermaseren, and A. Vogt. The three-loop splitting functions in QCD: The non-singlet case. *Nucl. Phys.*, B688:101–134, 2004.

[N$^+$96] T. Nicholls et al. Performance of an electromagnetic lead / scintillating fiber calorimeter for the H1 detector. *Nucl. Instrum. Meth.*, A374:149–156, 1996.

[Nau] J. Naumann. Entwicklung und Test der dritten H1-Triggerstufe. Univ. Dortmund, Germany (2003).

[P$^+$00] D. Pitzl et al. The H1 silicon vertex detector. *Nucl. Instrum. Meth.*, A454:334–349, 2000.

[P$^+$02] J. Pumplin et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 07:012, 2002.

[PB93] H. Plothow-Besch. PDFLIB: A Library of all available parton density functions of the nucleon, the pion and the photon and the corresponding alpha-s calculations. *Comput. Phys. Commun.*, 75:396–416, 1993.

[Pet10] Alexey Petrukhin. Measurements of the proton structure at HERA and their impact for LHC. *Nucl. Phys. Proc. Suppl.*, 207-208:45–48, 2010.

[S$^+$01] Torbjorn Sjostrand et al. High-energy physics event generation with PYTHIA 6.1. *Comput. Phys. Commun.*, 135:238–259, 2001.

[Scha] Stefan Schmitt. Regularized unfolding with a constraint.

[Schb] Stefan Schmitt. *TUnfold und TUnfoldSys, ROOT Packages available in ROOT Version 5.26 and later, see ROOT C++ class reference for documentation.*

[Sch96] S. Schleif. Spacal reconstruction. *H1 Software Note*, 56-03/96, 1996.

[SMS08] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[SP] Stefan Schmitt and Alexei Petrukhin. Qed compton lumi analysis, https://www-h1.desy.de/icgi-h1wiki/moin.cgi/elaninclusive/qed-compton-lumi.

[Spi92] H. Spiesberger. Monte Carlo treatment of radiative corrections at HERA. *Nucl. Phys. Proc. Suppl.*, 29A:221–228, 1992.

[Spi93] H. Spiesberger. Calculation and application of electroweak radiative corrections at HERA. *J. Phys.*, G19:1469–1478, 1993.

[SS] G. A. Schuler and H. Spiesberger. DJANGO: The Interface for the event generators HERACLES and LEPTO. In *Hamburg 1991, Proceedings, Physics at HERA, vol. 3* 1419-1432. (see HIGH ENERGY PHYSICS INDEX 30 (1992) No. 12988).

[Sti01] J. Stiewe. Hera (and non-hera) kinematics for pedestrians. *Universität Heidelberg*, 2001.

[VB98] E. Lohrmann V. Blobel. *Statistische und numerische Methoden der Datenanalyse.* Teubner Studienbücher, 1998.

[VMV04] A. Vogt, S. Moch, and J. A. M. Vermaseren. The three-loop splitting functions in QCD: The singlet case. *Nucl. Phys.*, B691:129–181, 2004.

[VVM05] J. A. M. Vermaseren, A. Vogt, and S. Moch. The third-order QCD corrections to deep-inelastic scattering by photon exchange. *Nucl. Phys.*, B724:3–182, 2005.

[WBG05] M. R. Whalley, D. Bourilkov, and R. C. Group. The Les Houches Accord PDFs (LHAPDF) and Lhaglue. 2005.

[Wei05a] Steven Weinberg. *The Quantum Theory of Fields: Volume 1, Foundations.* Cambridge University Press, 2005.

[Wei05b] Steven Weinberg. *The Quantum Theory of Fields: Volume 2, Modern Applications.* Cambridge University Press, 2005.

[Wei05c] Steven Weinberg. *The Quantum Theory of Fields: Volume 3, Supersymmetry.* Cambridge University Press, 2005.

[Zec] Gunter Zech. Comparing statistical data to Monte Carlo simulation: Parameter fitting and unfolding. DESY-95-113.

Danke!

Ich danke meinem Doktorvater Prof. Dr. Eckhard Elsen für die geduldige Betreuung dieser Arbeit sowie dafür, mich auf das spannende Thema Entfaltung gebracht zu haben. Er hat aus einem Theoretiker einen Fan experimenteller Messfehler gemacht. Ebensolcher Dank geht an Dr. Daniel Pitzl, der mir immer wieder viel wertvolle Zeit gewidmet hat und mich dabei von seinem enormen Detektorwissen hat profitieren lassen. Ferner danke ich Prof. Dr. Volker Blobel und Dr. Stefan Schmitt für zahlreiche hilfreiche Diskussionen über die theoretischen Aspekte des Entfaltens.

Ein grosses Dankeschön geht an das gesamte H1 Team für die anregende und herzliche Arbeitsatmosphäre. Das betrifft insbesondere meinen Freund und Kollegen Philipp Pahl, mit dem ich über drei Jahre hinweg das Büro geteilt habe.

Ganz besonders möchte ich meiner Frau Julia danken! Sie hat so manches Hoch und Tief im Entstehungsprozess dieser Arbeit liebevoll begleitet. Sie hat mir stets den Rücken gestärkt und an den Erfolg der Arbeit geglaubt, auch dann, wenn ich das zuweilen selbst nicht konnte. Für all deine Hilfe, liebe Julia, ganz lieben Dank!