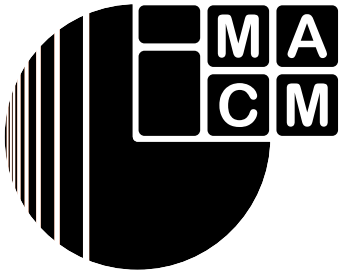

Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 11/04

Roland Pulch

Generalised Polynomial Chaos for Semi-Explicit Differential Algebraic Equations of Index 1

April 2011

http://www.math.uni-wuppertal.de

Generalised Polynomial Chaos for Semi-Explicit Differential Algebraic Equations of Index 1

Roland Pulch

Bergische Universität Wuppertal, Fachbereich Mathematik und Naturwissenschaften, Lehrstuhl für Angewandte Mathematik und Numerische Mathematik, Gaußstr. 20, D-42119 Wuppertal, Germany.
Email: pulch@math.uni-wuppertal.de

Abstract

Mathematical modelling of technical applications often yields systems of differential algebraic equations. Uncertainties of physical parameters can be considered by the introduction of random variables. A corresponding uncertainty quantification requires to solve the stochastic model. We focus on semi-explicit systems of nonlinear differential algebraic equations with index 1. The stochastic model is solved using the expansion of the generalised polynomial chaos. We investigate both the stochastic collocation technique and the stochastic Galerkin method to determine the unknown coefficient functions. In particular, we analyse the index of the larger coupled systems, which result from the stochastic Galerkin method. Numerical simulations of test examples are presented, where the two approaches are compared with respect to their efficiency.

1 Introduction

In technical applications, mathematical modelling of dynamical systems often results in differential algebraic equations (DAEs), i.e., a mixture of ordinary differential equations (ODEs) and algebraic equations. For example, network approaches yield large systems of DAEs corresponding to mechanical multibody dynamics or electric circuits, see [4, 6, 8]. Systems of DAEs exhibit qualitatively different properties than systems of ODEs. The index, which represents an integer number, indicates the level of the differences between a particular system of

DAEs and a general system of ODEs. Several concepts for the definition of an index exist. Numerical methods for initial value problems of ODEs are transferred into integrators for DAEs, where attention has to be paid also in dependence on the index of the systems, see [2, 7].

We assume that uncertainties are inherent in some physical parameters of the dynamical system. Corresponding parameters are replaced by random variables to achieve an uncertainty quantification. The random-dependent system of DAEs can be resolved by a quasi Monte-Carlo simulation, for example. Alternatively, we consider techniques based on the expansions of the generalised polynomial chaos (gPC), see [1, 5, 19]. The unknown coefficient functions can be determined either via a stochastic collocation method or via the stochastic Galerkin approach, see [17, 18]. Thereby, the stochastic Galerkin technique yields a larger coupled system of DAEs satisfied by an approximation of the coefficient functions.

The gPC expansions have already been applied for the simulation of systems of DAEs with random parameters in [11, 12, 13], where the focus is on periodic boundary value problems. In case of linear systems of DAEs, the index of the coupled systems of the stochastic Galerkin method is analysed in [14]. All index concepts are equivalent for linear DAEs. In this article, we consider semi-explicit systems of nonlinear DAEs with a differential index 1. For semi-explicit DAEs, the differential index is 1 if and only if the perturbation index is 1, see [3]. It is obvious that the index of the DAEs coincides within stochastic collocation methods. We analyse the index of the larger coupled system of DAEs, which is obtained by the stochastic Galerkin technique.

The approach of the stochastic collocation and the stochastic Galerkin method are compared in this paper. On the one hand, the properties of the involved systems of DAEs are analysed using the corresponding index. On the other hand, numerical simulations of initial value problems are performed to investigate the efficiency of each method, which is done by a comparison of both accuracy and computational effort.

The paper is organised as follows. We introduce the random-dependent systems of DAEs in Sect. 2. The stochastic collocation techniques and the stochastic Galerkin method are outlined. We analyse the index of the coupled systems from the Galerkin approach in Sect. 3. Numerical simulations of two test examples are presented in Sect. 4, where the efficiency of both gPC techniques is compared.

2 Stochastic Modelling

In this section, we define the stochastic model and apply the expansions of the generalised polynomial chaos for the corresponding solutions.

2.1 Problem Definition

We consider dynamical systems of the form

$$A(\mathbf{p})\dot{\mathbf{x}}(t,\mathbf{p}) = \mathbf{f}(t,\mathbf{x}(t,\mathbf{p}),\mathbf{p}), \tag{1}$$

where parameters $\mathbf{p} = (p_1,\dots,p_Q)$ with $\mathbf{p} \in \Pi \subseteq \mathbb{R}^Q$ are involved. The solution $\mathbf{x} : [t_0,t_1] \times \Pi \to \mathbb{R}^N$ depends on time as well as the parameters. In case of a regular mass matrix ($\det(A(\mathbf{p})) \neq 0$), the system (1) represents implicit ordinary differential equations (ODEs). In case of a singular mass matrix ($\det(A(\mathbf{p})) = 0$), we obtain a system (1) of differential algebraic equations (DAEs). We consider initial value problems

$$\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}), \tag{2}$$

where the initial values are allowed to depend on the parameters.

We assume that the parameters include some uncertainties. Consequently, we substitute the parameters by random variables

$$\mathbf{p} : \Omega \to \Pi, \quad \mathbf{p}(\omega) = (p_1(\omega),\dots,p_Q(\omega))$$

defined on some probability space $(\Omega,\mathcal{A},\mu)$. We apply independent distributions in this modelling, where a corresponding probability density function $\rho : \Pi \to \mathbb{R}$ is available. A random variable p can describe the perturbation of a physical parameter r, i.e.,

$$r(\omega) := \lambda p(\omega) + r_0 \tag{3}$$

with constants $\lambda, r_0 \in \mathbb{R}$. Thus a standardised variable with $\langle p\rangle = 0$ and $\langle p^2\rangle = 1$ can be used in the modelling, whereas the information on λ, r_0 is included in the system (1). The solution $\mathbf{x}(t,\mathbf{p})$ of the dynamical system (1) becomes a random process depending on time as well as the random parameters. We are interested in the key data of this random process like the expected value and the standard deviation, for example. More sophisticated data may also be resolved.

We define the function spaces

$$L^k(\Pi,\rho) := \left\{ f : \Pi \to \mathbb{R} : \int_\Pi |f(\mathbf{p})|^k \rho(\mathbf{p})\, \mathrm{d}\mathbf{p} < \infty \right\}$$

for each integer k. Given a function $f \in L^1(\Pi, \rho)$, we apply the notation

$$\langle f(\mathbf{p}) \rangle := \int_\Pi f(\mathbf{p}) \rho(\mathbf{p}) \, \mathrm{d}\mathbf{p} \tag{4}$$

for the corresponding expected value. For two functions $f, g \in L^2(\Pi, \rho)$, the expected value (4) implies the inner product

$$\langle f(\mathbf{p}) g(\mathbf{p}) \rangle = \int_\Pi f(\mathbf{p}) g(\mathbf{p}) \rho(\mathbf{p}) \, \mathrm{d}\mathbf{p}. \tag{5}$$

We employ this notation also to vector-valued functions and matrix-valued functions by each component separately.

In this paper, we consider semi-explicit systems of DAEs, i.e., the dynamical system (1) becomes

$$\begin{aligned} \mathbf{y}'(t, \mathbf{p}) &= \mathbf{f}(t, \mathbf{y}(t, \mathbf{p}), \mathbf{z}(t, \mathbf{p}), \mathbf{p}) \\ \mathbf{0} &= \mathbf{g}(t, \mathbf{y}(t, \mathbf{p}), \mathbf{z}(t, \mathbf{p}), \mathbf{p}) \end{aligned} \tag{6}$$

with the differential variables $\mathbf{y} : [t_0, t_1] \times \Pi \to \mathbb{R}^{N_y}$ and the algebraic variables $\mathbf{z} : [t_0, t_1] \times \Pi \to \mathbb{R}^{N_z}$. The right-hand sides exhibit the dimensions $\mathbf{f} \in \mathbb{R}^{N_y}$ and $\mathbf{g} \in \mathbb{R}^{N_z}$. We assume that the right-hand sides are continuous or sufficiently smooth if required.

A system of DAEs features different properties than a system of ODEs. The level of these differences is characterised by the index of the system of DAEs, where several index concepts exist, see [7]. We focus on semi-explicit systems (6) of differential index 1 and perturbation index 1. In case of semi-explicit DAEs, the differential index is 1 if and only if the perturbation index is 1. An equivalent condition is that the Jacobian matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}} \in \mathbb{R}^{N_z \times N_z}$ is regular, i.e.,

$$\det \left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right) \neq 0 \tag{7}$$

for all involved solutions and parameters in each time point $t \in [t_0, t_1]$. If (7) holds, then DAEs of the form (6) are also called Hessenberg DAEs of index 1. Dynamical systems (1) with a constant mass matrix ($A(\mathbf{p}) \equiv A_0$) can be transformed directly into an equivalent semi-explicit system (6) of the same dimension ($N_y + N_z = N$) and the same index. Moreover, each system of the form (1) can be converted into a corresponding semi-explicit system (6) with $N_y = N_z = N$ using $\mathbf{y} := \mathbf{x}$ and $\mathbf{z} := \mathbf{y}'$, where a higher index appears in general.

We specify an initial value problem via

$$\mathbf{y}(t_0, \mathbf{p}) = \mathbf{y}_0(\mathbf{p}), \quad \mathbf{z}(t_0, \mathbf{p}) = \mathbf{z}_0(\mathbf{p}) \tag{8}$$

with predetermined parameter-dependent functions $\mathbf{y}_0, \mathbf{z}_0$. In case of semi-explicit systems (6) of index 1, the initial values (8) must satisfy the consistency condition

$$\mathbf{g}(t_0, \mathbf{y}_0(\mathbf{p}), \mathbf{z}_0(\mathbf{p}), \mathbf{p}) = \mathbf{0} \tag{9}$$

for each $\mathbf{p} \in \Pi$. Hence the initial values $\mathbf{z}_0$ follow from the choice of the initial values $\mathbf{y}_0$ by the implicit function theorem. Even if the initial values $\mathbf{y}_0$ are independent of the parameters, the initial values $\mathbf{z}_0$ depend on the parameters if the function $\mathbf{g}$ does. Thus we consider parameter-dependent initial values (8) in general. For semi-explicit DAEs of higher index, hidden consistency conditions exist in addition to the algebraic constraints (9).

2.2 Generalised Polynomial Chaos

Considering random parameters, the stochastic model (6),(8) can be solved by a (quasi) Monte-Carlo simulation, for example. Alternatively, we consider spectral methods based on the polynomial chaos, see [5, 19]. We assume finite second moments of the components of the differential and algebraic variables corresponding to a solution of the stochastic model (6),(8). It follows that the expansions

$$\mathbf{y}(t, \mathbf{p}(\omega)) = \sum_{i=0}^{\infty} \mathbf{v}_i(t)\Phi_i(\mathbf{p}(\omega)), \quad \mathbf{z}(t, \mathbf{p}(\omega)) = \sum_{i=0}^{\infty} \mathbf{w}_i(t)\Phi_i(\mathbf{p}(\omega)) \tag{10}$$

converge with respect to the norm of $L^2(\Pi, \rho)$ for each $t \in [t_0, t_1]$. The series include orthogonal basis polynomials $\Phi_i : \Pi \to \mathbb{R}$. Thus let $\langle \Phi_i \Phi_j \rangle = \delta_{ij}$ with the Kronecker delta symbol. The basis polynomials follow from the probability distributions of the random parameters, see [16]. Thereby, the multivariate basis polynomials are just the products of corresponding univariate basis polynomials. If all random parameters exhibit Gaussian distributions, then the traditional homogeneous polynomial chaos appears. In case of non-Gaussian random parameters, we obtain the generalised polynomial chaos (gPC).

The coefficient functions $\mathbf{v}_i : [t_0, t_1] \to \mathbb{R}^{N_y}$ and $\mathbf{w}_i : [t_0, t_1] \to \mathbb{R}^{N_z}$ are unknown a priori. These time-dependent functions satisfy the equations

$$\mathbf{v}_i(t) = \langle \mathbf{y}(t, \mathbf{p})\Phi_i(\mathbf{p}) \rangle, \quad \mathbf{w}_i(t) = \langle \mathbf{z}(t, \mathbf{p})\Phi_i(\mathbf{p}) \rangle. \tag{11}$$

Assuming $\Phi_0 \equiv 1$, it follows $\mathbf{v}_0 = \langle \mathbf{y} \rangle$ and $\mathbf{w}_0 = \langle \mathbf{z} \rangle$.

In practice, the gPC expansions (10) have to be truncated. The resulting finite approximations read

$$\mathbf{y}^{(M)}(t, \mathbf{p}) := \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \quad \mathbf{z}^{(M)}(t, \mathbf{p}) := \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}) \tag{12}$$

for some integer M. Often all basis polynomials up to a certain degree are chosen in the finite sums. The coefficients in (12) yield approximations of the expected value and the variance of the random process. Nevertheless, more sophisticated quantities are also reproduced approximatively by this approach. For example, a truncated series (12) represents a surrogate model, which can be used to compute failure probabilities, cf. [9, 12].

2.3 Stochastic Collocation Techniques

We want to determine approximations of the coefficient functions involved in the truncated gPC expansion (12). Due to the property (11), the unknown coefficient functions represent evaluations of probabilistic integrals. Thus we achieve an approximation of the coefficient functions by a quadrature formula. We choose grid points $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(K)} \in \Pi$ in the domain of the parameters. It follows the approximations

$$\mathbf{v}_i(t) \doteq \sum_{k=1}^{K} \omega_k \Phi_i(\mathbf{p}^{(k)}) \mathbf{y}(t, \mathbf{p}^{(k)}), \quad \mathbf{w}_i(t) \doteq \sum_{k=1}^{K} \omega_k \Phi_i(\mathbf{p}^{(k)}) \mathbf{z}(t, \mathbf{p}^{(k)}). \tag{13}$$

For small numbers Q of parameters, a multivariate Gaussian quadrature can be employed straightforward, because the grids are tensor products of the nodes of the corresponding univariate Gaussian quadratures. For medium sized Q, sparse grids should be preferred. In case of large numbers Q of parameters, Monte-Carlo simulations with pseudo random numbers or quasi Monte-Carlo methods are applied. Examples for two random parameters ($Q = 2$) with independent standardised Gaussian distributions are shown in Figure 1.

Each technique of the form (13) is called a stochastic collocation, see [10, 17, 18]. The nodes $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(K)}$ can be seen as collocation points. In each method of this type, we have to solve K initial value problems (6),(8) of the original systems of DAEs. Thereby, the numerical methods constructed for the deterministic initial value problems of the DAEs are applied directly. Hence the stochastic collocation approach is also called the non-intrusive method.

2.4 Stochastic Galerkin Method

Inserting the truncated gPC expansions (12) into the semi-explicit system (6) yields the residuals

$$\begin{aligned} \mathbf{r}_y(t, \mathbf{p}) &:= \mathbf{y}^{(M)\prime}(t, \mathbf{p}) - \mathbf{f}(t, \mathbf{y}^{(M)}(t, \mathbf{p}), \mathbf{z}^{(M)}(t, \mathbf{p}), \mathbf{p}), \\ \mathbf{r}_z(t, \mathbf{p}) &:= \mathbf{g}(t, \mathbf{y}^{(M)}(t, \mathbf{p}), \mathbf{z}^{(M)}(t, \mathbf{p}), \mathbf{p}). \end{aligned} \tag{14}$$

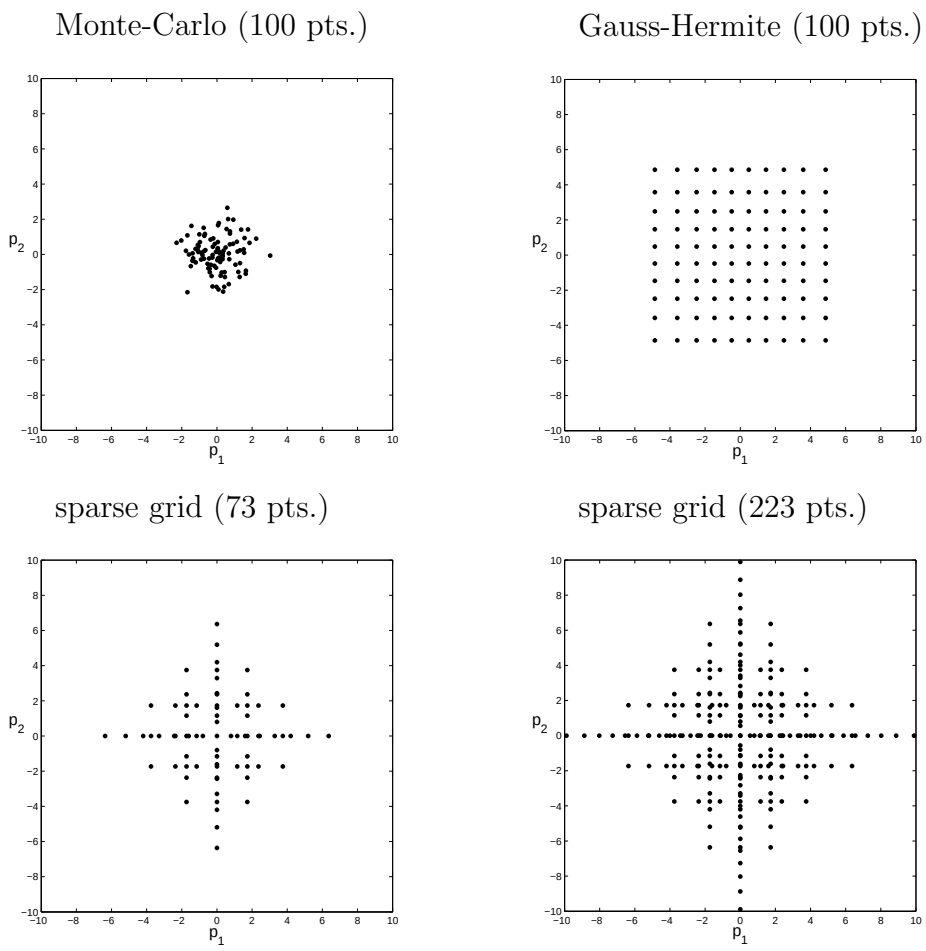

Figure 1: Examples for grids in stochastic collocation techniques with two independent Gaussian random variables.

We want to determine the coefficient functions such that the residuals become small in some sense. The Galerkin method requires the residuals to be orthogonal with respect to the space of the applied basis polynomials, i.e.,

$$\langle \mathbf{r}_{y,z}(t,\mathbf{p})\Phi_l(\mathbf{p})\rangle = \mathbf{0} \tag{15}$$

for $l = 0, 1, \ldots, M$ and each $t \in [t_0, t_1]$. Inserting the residuals (14) into the inner products (15), basic calculations lead to a larger coupled system, where the unknowns represent an approximation of the coefficient functions.

Definition 1 *The coupled system of the stochastic Galerkin method corresponding to the semi-explicit systems (6) reads*

$$\mathbf{v}_l'(t) = \left\langle \Phi_l(\mathbf{p}) \cdot \mathbf{f}\left(t, \sum_{i=0}^{M}\mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M}\mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle \tag{16}$$

$$\mathbf{0} = \left\langle \Phi_l(\mathbf{p}) \cdot \mathbf{g}\left(t, \sum_{i=0}^{M}\mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M}\mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle \tag{17}$$

for $l = 0, 1, \ldots, M$.

Although an exact solution of (16),(17) is not identical to the exact coefficients in (10), we apply the same symbols for convenience. The coupled system (16),(17) represents a semi-explicit system of DAEs again due to the orthogonality of the basis polynomials.

To obtain initial values for the coupled system (16),(17), the original initial values (8) can be expanded in the gPC. It follows

$$\mathbf{v}_l(t_0) = \langle \mathbf{y}_0(\mathbf{p})\Phi_l(\mathbf{p})\rangle, \quad \mathbf{w}_l(t_0) = \langle \mathbf{z}_0(\mathbf{p})\Phi_l(\mathbf{p})\rangle \tag{18}$$

for $l = 0, 1, \ldots, M$. However, since approximations of the type (12) are used, it holds $\mathbf{g} \neq \mathbf{0}$ at the solution of the coupled system in general. Hence the algebraic constraints (17) are not satisfied exactly, i.e., the straightforward choice (18) of the initial values is inconsistent. Alternatively, just the differential variables $\mathbf{v}_l(t_0)$ are computed via (18). We determine the algebraic variables $\mathbf{w}_l(t_0)$ by solving the $(M+1)N_z$ algebraic equations (17). Due to this construction, the initial values are consistent provided that the semi-explicit system (16),(17) exhibits also the index 1.

Solving the coupled system (16),(17) is also called the intrusive method. Given the original semi-explicit systems (6) of some index $I \geq 1$ for all parameters, we can solve the coupled system (16),(17) using the same numerical methods provided that the coupled system inherits the index I. If the coupled system exhibits a larger index, then disadvantages appear within the numerical simulation in comparison to the solution of the original systems.

3 Index Analysis

Let $S := \operatorname{supp}(\rho) = \overline{\{\mathbf{p} : \rho(\mathbf{p}) \neq 0\}} \subseteq \Pi \subseteq \mathbb{R}^Q$. We assume that a unique solution of the initial value problem (6),(8) exists for each $\mathbf{p} \in S$. Consequently, the Jacobian matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ depends just on the choice of $\mathbf{p}$ for fixed $t \in [t_0, t_1]$. The analysis of the index can be done for each t separately. The index-1 property is achieved for all $t \in [t_0, t_1]$ provided that the criteria hold for each $t \in [t_0, t_1]$. Hence we consider just a fixed $t \in [t_0, t_1]$ in the following.

3.1 Jacobian Matrix in Coupled System

For the algebraic part (17) of the coupled system from the stochastic Galerkin method, we use the abbreviation

$$\mathbf{G}_l := \left\langle \Phi_l(\mathbf{p}) \cdot \mathbf{g}\left(t, \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle = \mathbf{0}$$

for $l = 0, 1, \ldots, M$. Consequently, the Jacobian matrix, which determines the index-1 property of the coupled system, reads

$$\mathcal{G} := \left(\frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k}\right)_{l,k} \in \mathbb{R}^{(M+1)N_z \times (M+1)N_z}. \tag{19}$$

This matrix consists of the minors

$$\frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k} = \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \frac{\partial \mathbf{g}}{\partial \mathbf{z}}\left(t, \sum_{i=0}^{M} \mathbf{v}_i(t)\Phi_i(\mathbf{p}), \sum_{i=0}^{M} \mathbf{w}_i(t)\Phi_i(\mathbf{p}), \mathbf{p}\right)\right\rangle \tag{20}$$

for $l, k = 0, 1, \ldots, M$. The derivation of the formula (20) is based on the assumption that the differentiation and the probabilistic integration can be interchanged. Thus we assume that the entries of the Jacobian matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ are continuous in the closed domain S. Let the probability density function ρ also be continuous in S. If S is bounded, then it follows that the differentiation and the integration can be interchanged. In case of unbounded S, further integrability conditions are required to guarantee that formula (20) holds.

The condition $\det(\mathcal{G}) \neq 0$ is equivalent to the index-1 property of the coupled system (16),(17). In contrast, the coupled system exhibits an index at least 2 for $\det(\mathcal{G}) = 0$. An increase of the index represents a crucial drawback. Initial value problems of index 1 are well-posed with respect to the dependence on perturbed data. Initial value problems of index larger than 1 are, strictly speaking, ill-posed

with respect to the dependence on perturbed data, since the time derivative of the perturbation enters the problem, see [7].

Now we ask if the coupled system (16),(17) is of index 1 provided that the original systems (6) exhibit the index 1. As a first minor result, we obtain the following conclusion.

Theorem 1 *If the matrix* $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ *does not depend on* $\mathbf{y}, \mathbf{z}, \mathbf{p}$ *and the semi-explicit system (6) is of index 1, then the coupled system (16),(17) inherits the index 1.*

Proof:

The differential variables and the algebraic variables depend on the parameters. Due to the assumptions, the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ does not exhibit a dependence on the parameters at all. Now the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ depends only on time. It follows

$$\left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p})\frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right\rangle = \left\langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p}) \right\rangle \frac{\partial \mathbf{g}}{\partial \mathbf{z}}.$$

Since the system of basis polynomials is orthonormal, we obtain

$$\mathcal{G} = I_{M+1} \otimes \frac{\partial \mathbf{g}}{\partial \mathbf{z}}$$

with the identity matrix I_{M+1} and the Kronecker product of matrices. Hence the matrix $\mathcal{G}$ is regular if and only if the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ is regular. □

Even if the algebraic constraints in (6) described by $\mathbf{g}$ include $\mathbf{y}, \mathbf{z}, \mathbf{p}$, the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ is independent of $\mathbf{y}, \mathbf{z}, \mathbf{p}$ in some cases, see the example in Sect. 4.2. However, the assumption of Theorem 1 is often not given in practice.

In the following examinations, we will assume one of the following two properties.

Condition 1 *The matrix* $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ *corresponding to the semi-explicit system (6) is regular for all* $\mathbf{p} \in S = \mathrm{supp}(\rho)$.

Condition 2 *The matrix* $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ *corresponding to the semi-explicit system (6) is regular for almost all* $\mathbf{p}$*, i.e., all* $\mathbf{p} \in R \subseteq S$ *for some measurable set* R *where* $S \backslash R$ *has probability zero.*

Condition 1 and Condition 2 represent the index-1 property of the original systems for all parameters and almost all parameters, respectively.

In a stochastic collocation method, Condition 1 guarantees that all involved semi-explicit systems (6) are of index 1. However, systems (6) of larger index may appear if just the Condition 2 is valid.

3.2 Counterexample

Neither Condition 1 nor Condition 2 is sufficient for the regularity of the matrix $\mathcal{G}$ of the coupled system (16),(17). A corresponding counterexample exists already in the case $N_z = 1$ and a single parameter $Q = 1$. This counterexample can be embedded straightforward into examples with $N_z > 1$ and/or $Q > 1$.

We define the algebraic part

$$g(t, \mathbf{y}, z, p) := p \cdot z + u(\mathbf{y})$$

with an arbitrary function $u : \mathbb{R}^{N_y} \to \mathbb{R}$. Since it holds

$$\frac{\partial g}{\partial z} = p,$$

a corresponding system (6) is of index 1 for $p \neq 0$ and of index at least 2 for $p = 0$. The matrix of the corresponding coupled system (16),(17) consists of the entries

$$\mathcal{G} = (\langle p\Phi_l(p)\Phi_k(p)\rangle)_{l,k=0,1,\ldots,M}.$$

Due to the orthogonality of the basis polynomials, the matrix $\mathcal{G}$ is tridiagonal. We choose a symmetric probability density $\rho(p)$ around the critical point $p = 0$. The Condition 2 is always satisfied in this case. For example, we can apply a Gaussian distribution with mean value $p = 0$, see Figure 2 (left). However, the diagonal entries of $\mathcal{G}$ become zero. It follows that the matrix $\mathcal{G}$ has the property

$$\det(\mathcal{G}) \begin{cases} = 0 & \text{for even } M, \\ \neq 0 & \text{for odd } M. \end{cases}$$

We recognise that there is no integer M_0 such that the coupled system is of index 1 for all $M > M_0$. Hence an improvement of the accuracy of the gPC by increasing M does not omit this behaviour.

Moreover, we can choose a uniform distribution corresponding to the domain $S = [-b, -a] \cup [a, b]$ for some $0 < a < b$, see Figure 2 (right). It follows a symmetric probability density function again. Now the stronger Condition 1 is satisfied. Yet the matrix $\mathcal{G}$ is singular for even M again. The critical point $\frac{\partial g}{\partial z} = p = 0$ is not within the support of the probability density function. However, this critical point is situated in the convex hull of the support. We will reconsider this quality in Sect. 3.5.

The above counterexample also indicates that problems with respect to the index may appear in a stochastic collocation if Condition 2 is satisfied but not Condition 1. We suppose a Gaussian distribution with mean value $p = 0$. If we apply a Gauss-Hermite quadrature in the stochastic collocation method, then

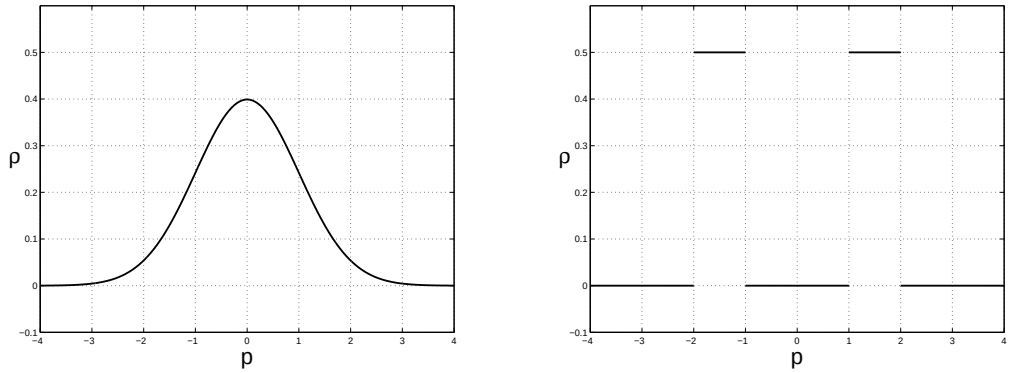

Figure 2: Special cases of probability density function ρ in counterexample: Gaussian distribution (left) and uniform distribution corresponding to the union of two disjoint intervals (right).

the critical point $p = 0$ is a node of this quadrature scheme in case of an odd number of nodes. It follows that a semi-explicit system (6) of index at least 2 has to be solved in the stochastic collocation, although almost all systems exhibit the index 1.

3.3 Limit Case of Small Variances

In this subsection, we apply a slightly different notation for the semi-explicit system (6) using physical parameters in the form (3). It follows the system

$$\begin{aligned} \mathbf{y}'_\lambda(t,\mathbf{p}) &= \mathbf{f}(t,\mathbf{y}_\lambda(t,\mathbf{p}),\mathbf{z}_\lambda(t,\mathbf{p}),\lambda\mathbf{p}+\mathbf{r}_0) \\ \mathbf{0} &= \mathbf{g}(t,\mathbf{y}_\lambda(t,\mathbf{p}),\mathbf{z}_\lambda(t,\mathbf{p}),\lambda\mathbf{p}+\mathbf{r}_0) \end{aligned} \tag{21}$$

depending on some $\lambda \in \mathbb{R}$. Similar modification have to be done within the initial values (8). For $\lambda = 0$, the system (21) becomes deterministic and involves the constant reference parameters $\mathbf{r}_0$. We assume that the index of the system of DAEs is 1 in case of the reference parameters. For $\lambda \neq 0$, a random perturbation appears. Using fixed random variables $\mathbf{p}$, the variance of the physical input parameters vanishes in the limit case $\lambda \to 0$.

The stochastic Galerkin method yields a corresponding coupled system with a matrix $\mathcal{G}_\lambda$ now, cf. (19). We obtain the following result in the limit of small variances, where the Kronecker product of matrices and the identity matrix I_{M+1} is involved. A subscript zero refers to the case $\lambda = 0$ and not to the initial values at t_0 now.

Theorem 2 *Assume that all functions within the matrix* $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ *are in* $L^2(\Pi, \rho)$ *and Lipschitz-continuous with respect to* $\mathbf{y}, \mathbf{z}$ *as well as the parameters. The matrix* $\mathcal{G}_\lambda$ *in the stochastic Galerkin method for the system (21) satisfies*

$$\lim_{\lambda \to 0} \mathcal{G}_\lambda = I_{M+1} \otimes \frac{\partial \mathbf{g}}{\partial \mathbf{z}}(t, \mathbf{y}_0(t, \mathbf{p}), \mathbf{z}_0(t, \mathbf{p}), \mathbf{r}_0) \tag{22}$$

provided that it holds

$$\lim_{\lambda \to 0} \left\langle \left\| \mathbf{y}_\lambda^{(M)}(t, \mathbf{p}) - \mathbf{y}_0(t, \mathbf{p}) \right\|^2 \right\rangle = 0, \quad \lim_{\lambda \to 0} \left\langle \left\| \mathbf{z}_\lambda^{(M)}(t, \mathbf{p}) - \mathbf{z}_0(t, \mathbf{p}) \right\|^2 \right\rangle = 0 \tag{23}$$

in an arbitrary vector norm $\|\cdot\|$.

Proof:

In the following, we apply the abbreviation

$$F(t) := \frac{\partial \mathbf{g}}{\partial \mathbf{z}}(t, \mathbf{y}_0(t, \mathbf{p}), \mathbf{z}_0(t, \mathbf{p}), \mathbf{r}_0).$$

We rearrange the minors (20) of the matrix $\mathcal{G}_\lambda$ to

$$\begin{aligned} \frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k} &= \langle \Phi_l(\mathbf{p}) \Phi_k(\mathbf{p}) F(t) \rangle \\ &\quad + \left\langle \Phi_l(\mathbf{p}) \Phi_k(\mathbf{p}) \left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}(t, \mathbf{y}_\lambda^{(M)}(t, \mathbf{p}), \mathbf{z}_\lambda^{(M)}(t, \mathbf{p}), \lambda \mathbf{p} + \mathbf{r}_0) - F(t) \right) \right\rangle. \end{aligned}$$

Using $\langle \Phi_l \Phi_k \rangle = \delta_{l,k}$, we write the complete matrix in the form

$$\mathcal{G}(t) = I_{M+1} \otimes F(t) + \mathcal{R}(t).$$

Let $D = (d_{ij}) \in \mathbb{R}^{N_z \times N_z}$ be the matrix consisting of the differences $\frac{\partial \mathbf{g}}{\partial \mathbf{z}} - F$, which is independent of l, k. The Cauchy-Schwarz inequality yields

$$|\langle \Phi_l \Phi_k d_{ij} \rangle| \leq \sqrt{\langle \Phi_l^2 \Phi_k^2 \rangle} \cdot \sqrt{\langle d_{ij}^2 \rangle} \leq \sqrt{\max_{i=0,1,\ldots,M} \langle \Phi_i^4 \rangle} \cdot \sqrt{\langle d_{ij}^2 \rangle}.$$

Without loss of generality, we apply the Euclidean vector norm $\|\cdot\|_2$ and the consistent Frobenius matrix norm $\|\cdot\|_*$. The Lipschitz-continuity of the functions in $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ allows for the conclusion

$$\|\langle \Phi_l \Phi_k D \rangle\|_*^2 \leq C_M \left\langle \left\| \mathbf{y}_\lambda^{(M)} - \mathbf{y}_0 \right\|_2^2 + \left\| \mathbf{z}_\lambda^{(M)} - \mathbf{z}_0 \right\|_2^2 + \|\lambda \mathbf{p}\|_2^2 \right\rangle$$

with a constant $C_M > 0$. Using the assumptions of the theorem, it follows

$$\lim_{\lambda \to 0} \|\langle \Phi_l \Phi_k D \rangle\|_* = 0$$

for all $l, k = 0, 1, \ldots, M$. Thus we obtain

$$\lim_{\lambda \to 0} \mathcal{R}(t) = 0,$$

which implies the formula (22). □

We motivate the assumption (23) further. It holds

$$\|\mathbf{y}_\lambda^{(M)}(t, \mathbf{p}) - \mathbf{y}_0(t, \mathbf{p})\| \leq \|\mathbf{y}_\lambda^{(M)}(t, \mathbf{p}) - \mathbf{y}_\lambda(t, \mathbf{p})\| + \|\mathbf{y}_\lambda(t, \mathbf{p}) - \mathbf{y}_0(t, \mathbf{p})\|.$$

The first term on the right-hand side represents the error of the stochastic Galerkin method. This error tends to zero in the norm $L^2(\Pi, \rho)$, see (23), for $M \to \infty$ provided that the stochastic Galerkin method is convergent. However, the constant C_M, which appears in the proof of Theorem 2, depends on M. Hence we do not achieve a condition uniformly for all $M \geq M_0$ with some sufficiently large M_0. The second term on the right-hand side converges to zero for $\lambda \to 0$ if the solution depends continuously on the physical parameters. The same discussion applies to the algebraic part $\mathbf{z}$.

Concerning the index of the corresponding systems of DAEs, we achieve the following result.

Corollary 1 *If the system (21) exhibits index 1 in case of the reference physical parameter* $\mathbf{r}_0$ *($\lambda = 0$), then the coupled system (16),(17) inherits the index 1 for sufficiently small variance of the input random variables provided that the assumptions of Theorem 2 are satisfied.*

The results of Theorem 2 and Corollary 1 can be generalised straightforward to the case of DAEs in Hessenberg form, see [7], with higher index, since the index is characterised by the regularity of specific matrices.

The above conclusions do not contradict the results from Sect. 3.2. The counterexample requires to choose a symmetric distribution around $p = 0$. Hence we have to select $r_0 = 0$ in this case, where the index-1 assumption is violated for the corresponding system (21).

The concept applied in this subsection implies only an asymptotic statement. We do not obtain a criterion on the index for fixed $\lambda \neq 0$. Hence further investigations are performed in the following subsections.

3.4 Dependence on Sign of Eigenvalues

A criterion for the index-1 property can be obtained by demanding that the signs of the eigenvalues of the Jacobian matrix do not change. We investigate the scalar

case ($N_z = 1$) first.

Theorem 3 *If it holds $\frac{\partial g}{\partial z} > 0$ for almost all $\mathbf{p}$ or $\frac{\partial g}{\partial z} < 0$ for almost all $\mathbf{p}$, then the matrix $\mathcal{G}$ from (19) is positive or negative definite, respectively.*

Proof:

Let $\mathbf{u} = (u_0, u_1, \ldots, u_M)^\top \in \mathbb{R}^{M+1}$ and $\mathbf{u} \neq 0$. It follows

$$\begin{aligned} \mathbf{u}^\top \mathcal{G} \mathbf{u} &= \sum_{l,k=0}^{M} u_l u_k \left\langle \Phi_l \Phi_k \frac{\partial g}{\partial z} \right\rangle = \left\langle \left(\sum_{l,k=0}^{M} u_l u_k \Phi_l \Phi_k \right) \frac{\partial g}{\partial z} \right\rangle \\ &= \left\langle \left(\sum_{l=0}^{M} u_l \Phi_l \right)^2 \frac{\partial g}{\partial z} \right\rangle. \end{aligned}$$

The latter probabilistic integral includes a non-negative polynomial. Due to $\mathbf{u} \neq 0$ and the linear independence of the basis polynomials, this polynomial is not identical to zero. Thus the number of zeros of the polynomial is finite. We obtain $\mathbf{u}^\top \mathcal{G} \mathbf{u} > 0$ for $\frac{\partial g}{\partial z} > 0$ and $\mathbf{u}^\top \mathcal{G} \mathbf{u} < 0$ for $\frac{\partial g}{\partial z} < 0$. □

In both cases of the theorem, the coupled system (16),(17) inherits the index-1 property from the original systems (6). Moreover, the property is independent of the choice of the subset of orthogonal basis polynomials. Theorem 3 does not contradict the results of Sect. 3.2. The counterexample applies a symmetric probability distribution around the critical point $\frac{\partial g}{\partial z} = p = 0$. Hence both positive and negative values appear and the assumptions of Theorem 3 are not satisfied.

The results of the scalar case can be generalised to the multidimensional case $N_z > 1$ under additional assumptions.

Theorem 4 *Let the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ be real diagonalisable in the form*

$$\frac{\partial \mathbf{g}}{\partial \mathbf{z}} = U(t) D(t, \mathbf{p}) U(t)^{-1} \tag{24}$$

with a regular matrix $U(t) \in \mathbb{R}^{N_z \times N_z}$ and a diagonal matrix $D(t, \mathbf{p})$ including the entries $\lambda_i(t, \mathbf{p})$ for $i = 1, \ldots, N_z$. If each eigenvalue λ_i is either positive or negative for almost all $\mathbf{p}$, then the matrix $\mathcal{G}$ from (19) is regular.

Proof:

The minors of the matrix $\mathcal{G}$ can be written as

$$\begin{aligned} \frac{\partial \mathbf{G}_l}{\partial \mathbf{w}_k} &= \left\langle \Phi_l(\mathbf{p}) \Phi_k(\mathbf{p}) \frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right\rangle = \left\langle \Phi_l(\mathbf{p}) \Phi_k(\mathbf{p}) U(t) D(t, \mathbf{p}) U(t)^{-1} \right\rangle \\ &= U(t) \left\langle \Phi_l(\mathbf{p}) \Phi_k(\mathbf{p}) D(t, \mathbf{p}) \right\rangle U(t)^{-1} \end{aligned}$$

for $l, k = 0, 1, \ldots, M$. We obtain the similarity transformation

$$\mathcal{G}(t) = (I_{M+1} \otimes U(t))\hat{\mathcal{G}}(t)(I_{M+1} \otimes U(t)^{-1}).$$

The matrix $\hat{\mathcal{G}}(t)$ consists of the minors

$$(\hat{\mathcal{G}}(t))_{l,k} = \langle \Phi_l(\mathbf{p})\Phi_k(\mathbf{p})D(t,\mathbf{p})\rangle \,.$$

Let $B(\mathbf{p}) := (\Phi_l(\mathbf{p})\Phi_k(\mathbf{p})) \in \mathbb{R}^{(M+1)\times(M+1)}$. It exists a specific permutation matrix $\mathcal{P} \in \mathbb{R}^{(M+1)N_z\times(M+1)N_z}$ independent of t such that the transformed matrix exhibits a block diagonal structure with the minors

$$(\mathcal{P}\hat{\mathcal{G}}(t)\mathcal{P})_{i,i} = \langle \lambda_i(t,\mathbf{p})B(\mathbf{p})\rangle$$

for $i = 1, \ldots, N_z$. The entries of each diagonal block are the same as in the matrix of the coupled system for the case $N_z = 1$ with λ_i instead of $\frac{\partial g}{\partial z}$. Theorem 3 implies that each diagonal block is positive or negative definite provided that the eigenvalues λ_i do not change their signs. Since the regularity of a matrix is invariant under permutations of rows and columns, it follows $\det(\hat{\mathcal{G}}(t)) \neq 0$ and thus $\det(\mathcal{G}(t)) \neq 0$. □

The assumptions of Theorem 4 are relatively strong. Firstly, the Jacobian matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ has to be real diagonalisable, which excludes matrices with complex eigenvalues, for example. Secondly, the transformation matrices U are assumed to be independent of the parameters $\mathbf{p}$. Nevertheless, both properties are given in case of a diagonal matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$.

Concerning the index-1 property, we obtain the following implication.

Corollary 2 *Let the assumptions of Theorem 4 be fulfilled. Moreover, let the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ depend continuously on the parameters. If the domain $S = supp(\rho)$ is path-connected, then Condition 1 implies an index of 1 for the coupled system (16),(17).*

Proof:

The assumption (24) yields

$$U(t)^{-1}\frac{\partial \mathbf{g}}{\partial \mathbf{z}}(t,\mathbf{p})U(t) = D(t,\mathbf{p}).$$

Hence the eigenvalues depend continuously on the parameters. Condition 1 implies that each eigenvalue is non-zero for all $\mathbf{p} \in S$. It follows that an eigenvalue does not change its sign. Otherwise, we obtain a path between two points in S, where an eigenvalue zero appears on the path. Now Theorem 4 yields the index-1 property. □

The counterexample of Sect. 3.2, where Condition 1 is satisfied, does not involve a path-connected domain S. Furthermore, a convex domain is always path-connected. We will retrieve this property in the next subsection.

3.5 Criterion from Numerical Range

For a matrix $C \in \mathbb{C}^{N\times N}$, the numerical range

$$W(C) := \left\{ \mathbf{u}^* C \mathbf{u} : \mathbf{u} \in \mathbb{C}^N,\ \mathbf{u}^*\mathbf{u} = 1 \right\} \subset \mathbb{C}$$

represents a closed and convex set. The spectrum of C is a subset of $W(C)$. We define the numerical range of our random Jacobian matrix as

$$\widehat{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) := \left\{ \mathbf{u}^* \tfrac{\partial \mathbf{g}}{\partial \mathbf{z}} \mathbf{u} : \mathbf{u} \in \mathbb{C}^{N_z},\ \mathbf{u}^*\mathbf{u} = 1,\ \mathbf{p} \in S \right\} = \bigcup_{\mathbf{p}\in S} W\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right). \tag{25}$$

This set is not necessarily closed or convex within $\mathbb{C}$. However, the numerical range (25) is larger than required for our purposes. We apply an essential numerical range introduced in [15].

Definition 2 *Let $B(z,\varepsilon)$ be the ball of radius ε centered at $z \in \mathbb{C}$ and*

$$A(z,\varepsilon) := \left\{ \mathbf{p} \in S : B(z,\varepsilon) \cap W(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}) \neq \emptyset \right\}.$$

The essential numerical range of the random Jacobian matrix is

$$\widetilde{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) := \left\{ z \in \mathbb{C} : \int_{A(z,\varepsilon)} \rho(\mathbf{p})\, \mathrm{d}\mathbf{p} > 0 \ \textit{ for all } \varepsilon > 0 \right\}.$$

In comparison to the numerical range (25), the definition of the essential numerical range yields

$$\widetilde{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right) \subseteq \overline{\widehat{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)}.$$

It follows

$$\operatorname{conv}\left(\widetilde{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)\right) \subseteq \operatorname{conv}\left(\overline{\widehat{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)}\right) \tag{26}$$

for the convex hull of the sets.

Theorem 5 *If $0 \notin \operatorname{conv}\left(\widetilde{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)\right)$ holds, then the matrix $\mathcal{G}$ from (19) is regular for an arbitrary choice of an orthogonal basis.*

Proof:

Theorem 2 in [15] implies that

$$\operatorname{spect}(\mathcal{G}) \subset \operatorname{conv}\left(\widetilde{W}\left(\tfrac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)\right)$$

holds for the spectrum of the gPC matrix $\mathcal{G}$. If $0 \notin \operatorname{conv}\left(\widetilde{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)\right)$ is satisfied, then it follows $0 \notin \operatorname{spect}(\mathcal{G})$ and the matrix $\mathcal{G}$ is regular. $\square$

The essential numerical range may be difficult to determine. Applying the inclusion (26), we obtain the following criterion, which is often easier to verify provided that it is fulfilled by the problem.

Corollary 3 *If* $0 \notin \operatorname{conv}\left(\overline{\widehat{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)}\right)$ *holds, then the matrix* $\mathcal{G}$ *from (19) is regular for an arbitrary choice of an orthogonal basis.*

We consider problems, where the original semi-explicit systems exhibit the index 1 for all $\mathbf{p} \in S$, see Condition 1. We conclude that zero is not in the spectrum of $\frac{\partial \mathbf{g}}{\partial \mathbf{z}}$ for all $\mathbf{p} \in S$. Yet this condition is not sufficient to guarantee $0 \notin \widehat{W}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}}\right)$ in the case $N_z > 1$.

For $N_z = 1$, we obtain more potential for conclusions. The numerical range (25) simplifies to

$$\widehat{W}\left(\tfrac{\partial g}{\partial z}\right) = \left\{\tfrac{\partial g}{\partial z} : \mathbf{p} \in S\right\} \subseteq \mathbb{R}. \tag{27}$$

Consequently, Condition 1 of the original semi-explicit systems (6) implies the property $0 \notin \widehat{W}\left(\frac{\partial g}{\partial z}\right)$. Due to Theorem 5, we have to consider the convex hull of the numerical range. However, the condition $0 \notin \operatorname{conv}\left(\overline{\widehat{W}\left(\frac{\partial g}{\partial z}\right)}\right)$ is equivalent to $\frac{\partial g}{\partial z} \geq \eta > 0$ or $\frac{\partial g}{\partial z} \leq \eta < 0$ provided that the partial derivative depends continuously on the parameters. Since the assumption can be weakened using the essential numerical range, the criterion becomes $\frac{\partial g}{\partial z} > 0$ or $\frac{\partial g}{\partial z} < 0$. Therefore we achieve the same conclusion as in Sect. 3.4.

The assumptions in Theorem 5 and Corollary 3 are relatively strong, since neither Condition 1 nor Condition 2 are sufficient to guarantee these requirements in general. Nevertheless, we have to demand this criterion due to the counterexample in Sect. 3.2, since it holds

$$0 \notin \overline{\widehat{W}\left(\tfrac{\partial g}{\partial z}\right)}, \qquad 0 \in \operatorname{conv}\left(\overline{\widehat{W}\left(\tfrac{\partial g}{\partial z}\right)}\right)$$

in case of a uniform distribution with $S = [-b, -a] \cup [a, b]$, for example.

4 Test Examples

We apply a stochastic collocation technique as well as the stochastic Galerkin method to two test problems now.

4.1 First Example

As a benchmark, we employ a simple test example with $N_y = N_z = 1$, i.e.,

$$\begin{aligned} y'(t) &= -(1 + 0.1p_1)y(t) + z(t)^3 \\ 0 &= -(1 + 0.1p_2)y(t) - z(t)^3 \end{aligned} \tag{28}$$

with two parameters p_1, p_2. It holds

$$\frac{\partial g}{\partial z} = -3z(t)^2 < 0 \quad \text{for all } \; z(t) \neq 0. \tag{29}$$

Thus the index of the system (28) is 1 for arbitrary parameters provided that $z \neq 0$. We specify the consistent initial values

$$y(0) = 1, \quad z(0) = -\sqrt[3]{1 + 0.1p_2}. \tag{30}$$

The exact solution of the initial value problem (28),(30) reads

$$y(t) = \mathrm{e}^{-(2+0.1p_1+0.1p_2)t}, \quad z(t) = -\sqrt[3]{(1 + 0.1p_2)y(t)}. \tag{31}$$

If we arrange distributions of the random parameters with a bounded domain S and sufficiently small values p_1, p_2, then the problem satisfies Condition 1. Alternatively, we choose two independent Gaussian random variables with mean zero and unit variance for p_1, p_2 now. Both increasing and decreasing exponential functions y appear, where the probability of a decreasing process is much higher. Although the problem fulfils Condition 2 only, Theorem 3 guarantees that the coupled system of the stochastic Galerkin approach exhibits index 1 due to the property (29).

Since Gaussian distributions are considered, the gPC applies the Hermite polynomials. We include all two-variate polynomials up to degree 3 in the truncated series (12), which results in ten basis polynomials ($M = 9$). On the one hand, a stochastic collocation yields approximations of the coefficient functions based on a two-dimensional Gauss-Hermite quadrature with a grid of size 7×7. On the other hand, the stochastic Galerkin method requires to solve the coupled system (16),(17), where the probabilistic integrals in the right-hand sides are discretised using a Gauss-Hermite quadrature on a grid of size 7×7 again. Since the right-hand side of (28) consists of polynomials, the evaluations of the probabilistic integrals are exact except for roundoff errors in the stochastic Galerkin technique.

We compute the solutions within the time interval $[t_0, t_1] = [0, 10]$. In each method, the backward differentiation formula (BDF) of second order solves the systems of DAEs, see [7]. Thereby, we apply a constant step size $\Delta t = \frac{1}{50}$.

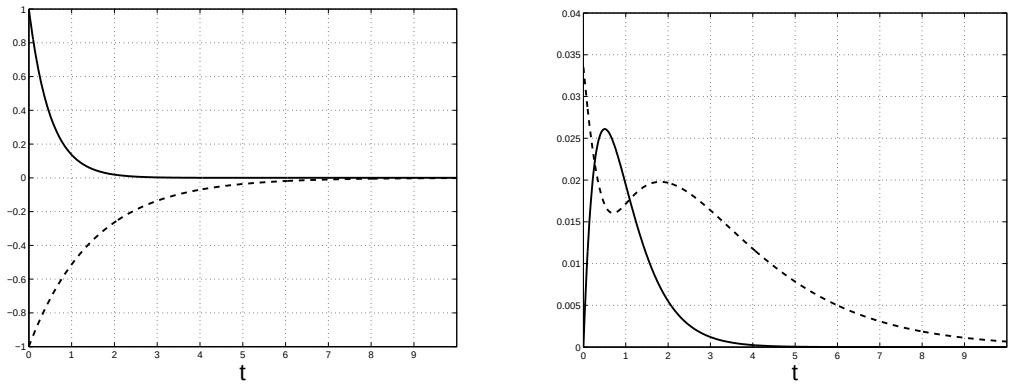

Figure 3: Expected values (left) and standard deviation (right) for differential component y (solid line) and algebraic component z (dashed line).

Since both time integration and discretisation in probability space employ the same schemes and grids, the number of right-hand side evaluations of the original systems (28) is identical in both stochastic collocation and stochastic Galerkin method. Just the linear algebra part of the Newton iterations within the implicit time integration causes a computational overhead in the stochastic Galerkin technique. The computations have been performed in the software package MATLAB. The CPU time of the collocation and the Galerkin method was 4.8 and 22.9, respectively.

We illustrate the results of the stochastic collocation. Figure 3 shows the expected values and the standard deviation of the stochastic processes. Furthermore, Figure 4 depicts the coefficient functions of the gPC expansions for higher degrees. Remark that some coefficient functions coincide for the differential component y. Table 1 illustrates the magnitude of the coefficient functions, where the maximum value is calculated on the grid in time. We observe an exponential decay of the coefficients, which is typical for the convergence of a gPC expansion in case of a sufficiently smooth dependence on the random parameters.

To compare the accuracy of the stochastic collocation and the stochastic Galerkin method, we compute a reference solution using the exact solutions (31) in a stochastic collocation with a two-dimensional Gauss-Hermite quadrature on a grid of size 14×14. Figure 5 visualises the maximum differences of the coefficient functions with respect to the reference solution. It follows that the accuracy coincides for the coefficients of lower degree, since the approximations of both methods are nearly the same due to the identical discretisation schemes. However, the stochastic collocation results in a better accuracy for coefficients of higher degree. This fact is surprising, since we omit numerical errors in the approximations of probabilistic integrals within the stochastic Galerkin method.

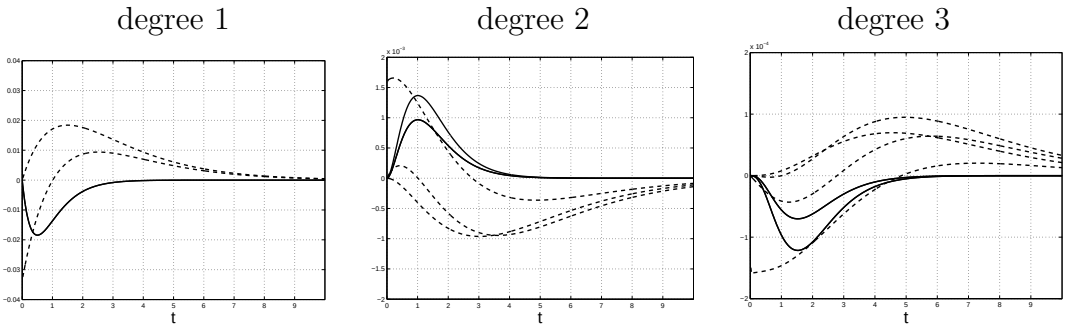

Figure 4: Coefficient functions in gPC expansion for differential component y (solid line) and algebraic component z (dashed line).

Table 1: Maximum values in time of coefficient functions in gPC expansion.

	diff. comp. y	alg. comp. z
degree 1	$1.8 \cdot 10^{-2}$	$3.2 \cdot 10^{-2}$
	$1.8 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$
degree 2	$9.7 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$
	$1.4 \cdot 10^{-3}$	$9.5 \cdot 10^{-4}$
	$9.7 \cdot 10^{-4}$	$9.6 \cdot 10^{-4}$
degree 3	$7.0 \cdot 10^{-5}$	$1.6 \cdot 10^{-4}$
	$1.2 \cdot 10^{-4}$	$6.4 \cdot 10^{-5}$
	$1.2 \cdot 10^{-4}$	$9.5 \cdot 10^{-5}$
	$7.0 \cdot 10^{-5}$	$7.0 \cdot 10^{-5}$

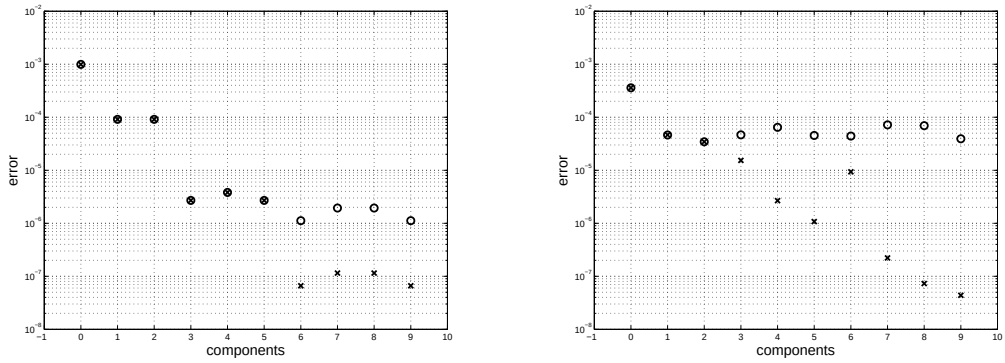

Figure 5: Maximum differences in reference solution vs. st. Galerkin (circles) and reference solution vs. st. collocation (crosses) for coefficient functions of differential component (left) and algebraic component (right) – semilog. scale.

The differences decrease exponentially in both techniques, since the magnitudes of the exact coefficient functions also decrease exponentially, cf. Table 1.

4.2 Second Example

Mathematical modelling of electric circuits typically results in systems of DAEs, see [6, 8]. We consider the electric circuit of a voltage controlled oscillator depicted in Figure 6. A particular modelling yields the semi-explicit system

$$\begin{aligned} u'(t) &= \tfrac{1}{C}\imath_C(t) \\ \imath_L'(t) &= \tfrac{1}{L}u(t) \\ 0 &= \imath_R(t) - (G_0 - G_\infty)U_0 \tanh\left(\tfrac{u(t)}{U_0}\right) - G_\infty u(t) \\ 0 &= \imath_C(t) + \imath_L(t) + \imath_R(t) \end{aligned} \tag{32}$$

for the unknown node voltage u and the branch currents $\imath_L, \imath_R, \imath_C$. The index of the system (32) is always equal to one. We arrange the physical parameters

$$C = 10^{-9}\ \mathrm{F},\ \ L = 10^{-6}\ \mathrm{H},\ \ U_0 = 1\ \mathrm{V},\ \ G_0 = -0.1\ \mathrm{A/V},\ \ G_\infty = 0.25\ \mathrm{A/V}.$$

The solutions of initial value problems tend to a periodic limit cycle. Therefore this example has been applied in [13] for the investigation of periodic boundary value problems. Alternatively, we consider the initial value problem

$$u(0) = 0\ \mathrm{V}, \quad \imath_L(0) = 0.1\ \mathrm{A}, \quad \imath_R(0) = 0\ \mathrm{A}, \quad \imath_C(0) = -0.1\ \mathrm{A},$$

which represents a consistent choice for arbitrary parameters. Numerical simulations are performed within the time interval $[t_0, t_1] = [0\mathrm{s}, 10^{-6}\mathrm{s}]$.

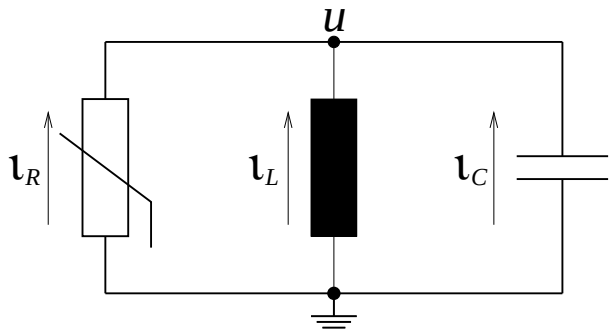

Figure 6: Electric circuit of voltage controlled oscillator.

In the stochastic modelling, we replace the two deterministic parameters C, G_∞ by the random parameters

$$\widetilde{C}(p_1) := C(1 + 0.01p_1), \quad \widetilde{G}_\infty(p_2) := G_\infty(1 + 0.01p_2)$$

with independent random variables p_1, p_2 both uniformly distributed in the interval $[-1, 1]$. Hence it holds $\widetilde{C}, \widetilde{G}_\infty > 0$ for all realisations of the random variables, which is necessary to achieve physically reasonable systems (32). Both parameters $\widetilde{C}$ and $\widetilde{G}_\infty$ influence the frequency of the corresponding solutions.

The gPC expansions are based on the Legendre polynomials now. We use all two-variate basis polynomials up to degree 3 in the truncated series (12) again ($M = 9$). Although the algebraic part of (32) includes a random parameter, this example satisfies the assumptions of Theorem 1. Hence the coupled system (16),(17) inherits the index-1 property in the stochastic Galerkin approach. The stochastic collocation technique and the stochastic Galerkin method are applied in the same form as in the previous example. Gauss-Legendre quadrature yields approximations of probabilistic integrals using a grid of size 4×4 now. Time integration is based on the BDF scheme of second order again with constant step size $\Delta t = 10^{-9}$s.

Figure 7 and Figure 8 illustrate the expected values and standard deviations of the components of the random process, which are reconstructed by the solutions from the stochastic collocation method. Since solutions of the original system (32) corresponding to different parameters exhibit different frequencies, the variances increase in time. However, the variance does not tend to infinity but to a periodic state indicating high uncertainties. Nevertheless, the order of the truncated gPC expansions (12) has to be increased for larger times to guarantee sufficiently accurate approximations, i.e., larger values M must be applied.

Using the software package MATLAB, the CPU times of the stochastic collocation and the stochastic Galerkin method were 3.2 and 40.64, respectively. To compare the accuracy, we compute a reference solution via stochastic collocation

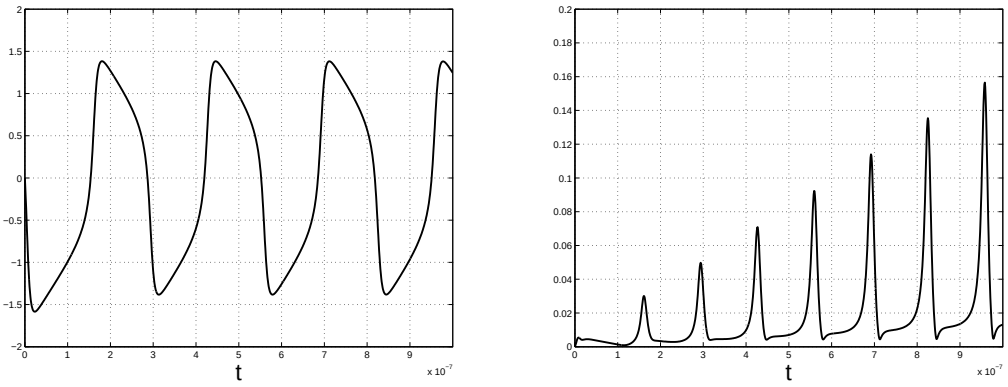

Figure 7: Expected values (left) and standard deviation (right) of node voltage u with unit [V] in oscillator.

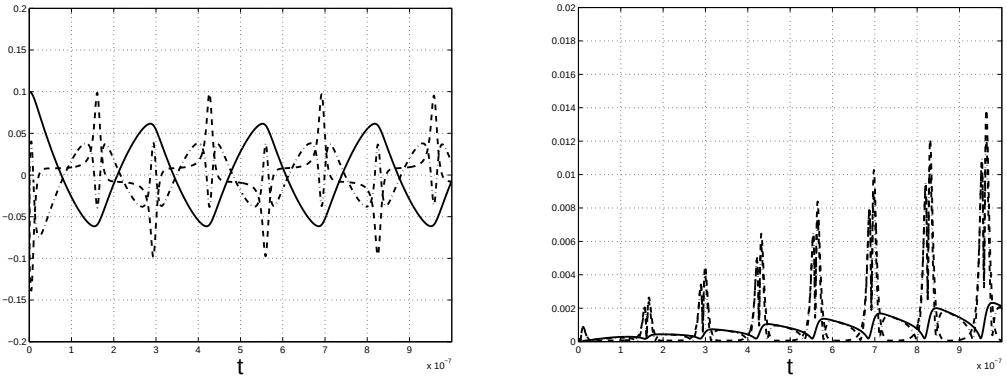

Figure 8: Expected values (left) and standard deviation (right) of branch currents $\imath_L$ (solid line), $\imath_C$ (dashed line), $\imath_R$ (dashed-dot line) with unit [A] in oscillator.

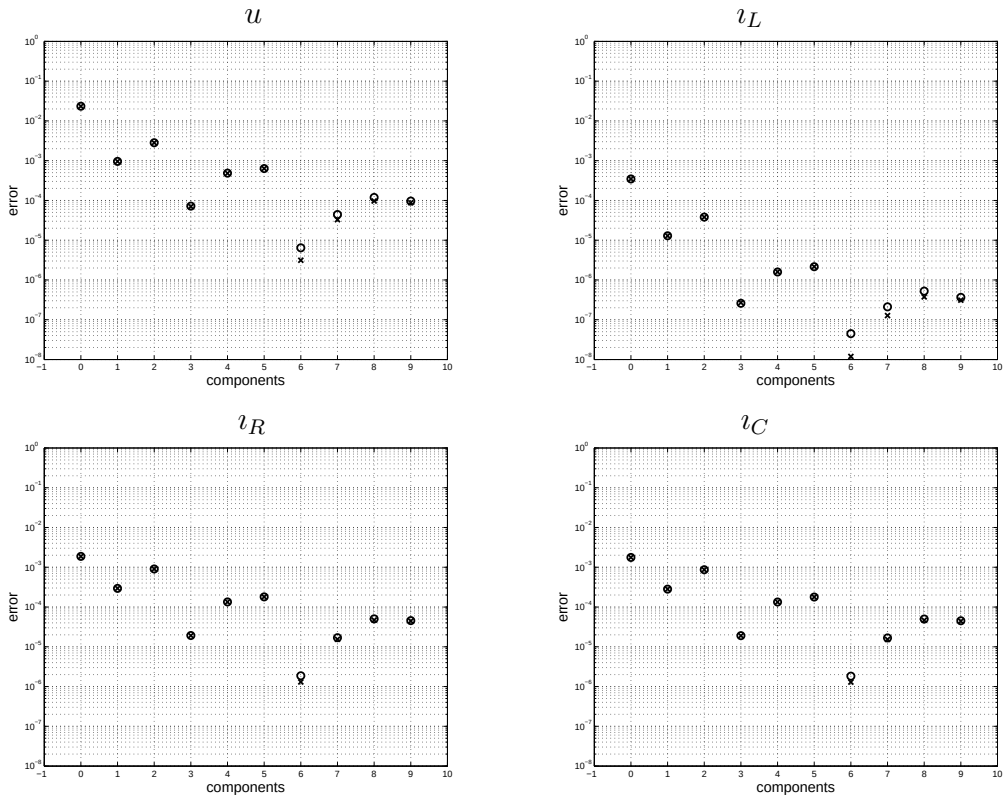

Figure 9: Maximum differences in reference solution vs. st. Galerkin (circles) and reference solution vs. st. collocation (crosses) for coefficient functions of gPC expansion for the four components of the random process – semilog. scale.

including Gauss-Legendre quadrature on a grid of size 8×8 and time integration with step size $\Delta t = 0.5 \cdot 10^{-9}$s. Figure 9 shows the maximum differences of the approximations with respect to the reference solution. We observe a good agreement of the achieved accuracies within the stochastic collocation and the stochastic Galerkin technique, which is caused by the application of the same discretisation schemes again. However, the computational effort of the stochastic Galerkin method is significantly larger due to the linear algebra part.

5 Conclusions

The approach of the generalised polynomial chaos has been applied to semi-explicit systems of differential algebraic equations with index 1, which include random parameters. Either a stochastic collocation technique or a stochastic Galerkin method can be used to compute the unknown coefficient functions of the

expansions. The presented results indicate that a stochastic collocation should be preferred over the stochastic Galerkin approach. On the one hand, the theoretical investigations show that the index of the larger coupled system from the Galerkin method can increase in comparison to the original systems of index 1. On the other hand, the numerical simulations illustrate that the Galerkin method requires a larger computational work due to the linear algebra part, whereas the accuracy is nearly the same in both techniques. Nevertheless, the differences with respect to the efficiency may become smaller in other types of problems. For example, the linear algebra part can be negligible if the evaluations of the functions in the original systems are expensive. We expect a similar behaviour in both the theoretical properties and the numerical simulations for general systems of differential algebraic equations with a possibly higher index.

References

[1] F. Augustin, A. Gilg, M. Paffrath, P. Rentrop, U. Wever: Polynomial chaos for the approximation of uncertainties: chances and limits. Euro. Jnl. of Applied Mathematics 19, pp. 149–190.

[2] K.E. Brenan, S.L. Campbell, L.R. Petzold: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. SIAM, Philadelphia, 1996.

[3] S. L. Campbell, C. W. Gear: The index of general nonlinear DAEs. Numer. Math. 72 (1995), pp. 173–196.

[4] E. Eich-Soellner, C. Führer: Numerical Methods in Multibody Dynamics. Teubner, Stuttgart, 1998.

[5] R.G. Ghanem, P. Spanos: Stochastic Finite Elements: A Spectral Approach. Springer, New York, 1991.

[6] M. Günther, U. Feldmann: CAD based electric circuit modeling in industry I: mathematical structure and index of network equations. Surv. Math. Ind. 8 (1999), pp. 97–129.

[7] E. Hairer, G. Wanner: Solving Ordinary Differential Equations. Vol. 2: Stiff and Differential-Algebraic Equations. 2nd ed., Springer, Berlin, 1996.

[8] W. Kampowsky, P. Rentrop, W. Schmitt: Classification and numerical simulation of electric circuits. Surv. Math. Ind. 2 (1992), pp. 23–65.

[9] J. Li, D. Xiu: Evaluation of failure probability via surrogate models. J. Comput. Phys. 229 (2010), pp. 8966–8980.

[10] A. Loeven, J. Witteveen, H. Bijl: Probabilistic collocation: An efficient non-intrusive approach for arbitrarily distributed parametric uncertainties. In: M.B. Bragg (ed.): Proceedings of 45th AIAA Aerospace Sciences Meeting and Exhibit 2007. pp. 1–14.

[11] R. Pulch: Polynomial chaos for analysing periodic processes of differential algebraic equations with random parameters. Proc. Appl. Math. Mech. 8 (2008), pp. 10069–10072.

[12] R. Pulch: Polynomial chaos for the computation of failure probabilities in periodic problems. In: J. Roos, L. Costa (eds.): Scientific Computing in Electrical Engineering. Mathematics in Industry Vol. 14, Springer, Berlin 2010, pp. 191–198.

[13] R. Pulch: Modelling and simulation of autonomous oscillators with random parameters. Math. Comput. Simulat. 81 (2011), pp. 1128–1143.

[14] R. Pulch: Polynomial chaos for linear differential algebraic equations with random parameters. to appear in: International Journal for Uncertainty Quantification.

[15] B. Sonday, R. Berry, B. Debusschere, H. Najm: Eigenvalues of the Jacobian of a Galerkin-projected uncertain ODE system. to appear in: SIAM Journal of Scientific Computing.

[16] D. Xiu, G.E. Karniadakis: The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. 24 (2002) 2, pp. 619–644.

[17] D. Xiu, J.S. Hesthaven: High order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. 27 (2005) 3, pp. 1118–1139.

[18] D. Xiu: Fast numerical methods for stochastic computations: a review. Comm. Comput. Phys. 5 (2009), pp. 242–272.

[19] D. Xiu: Numerical methods for stochastic computations: a spectral method approach. Princeton University Press, 2010.