Technische Universität München

Professur für Kontinuumsmechanik

Sparse Variational Bayesian algorithms for large-scale inverse problems with applications in biomechanics

Isabell Maria Franck

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor - Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr.-Ing. Wolfgang A. Wall
Prüfer der Dissertation:
Prof. Phaedon-Stelios Koutsourelakis, Ph.D.
Prof. Nicholas Zabaras, Ph.D., University of Notre Dame

Die Dissertation wurde am 23.01.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 15.03.2017 angenommen.

Abstract

Accurate uncertainty quantification for model-based, large-scale inverse problems represents one of the fundamental challenges in the context of computational science and engineering. To perform statistical inference for a high number of unknown variables, state-of-the-art sampling methods are intractable, since they require an exuberant number of expensive forward calls of the model. In order to address this problem, an efficient and accurate Bayesian framework is developed in this thesis. It is based on a Variational Bayesian formulation that aims at approximating the exact posterior density by solving an optimization problem over an appropriately selected family of distributions. This enables the computation of a good approximation of the exact posterior density with very few forward calls.

The main goals of this work are: to overcome the limitations of current algorithms which cope with the curse of dimensionality, and to accurately quantify uncertainties. Firstly, a new and effective technique for dimensionality reduction is proposed. It reveals low-dimensional subspaces where the variance of the approximate posterior density is concentrated. This technique is based on a novel dictionary learning method based on a fully Bayesian formulation. Secondly, the challenge of high accuracy in posterior inference is addressed, which is especially difficult to fulfill for multimodal posteriors. A novel Variational Bayesian strategy is developed that approximates the posterior using a mixture of multivariate Gaussian distributions. Each of these mixture components provides an accurate local approximation of the posterior by identifying a different, low-dimensional subspace. Finally, besides the presence of observational noise and parametric uncertainty, another source of uncertainty - up to the present day barely investigated - is quantified in this thesis. More precisely, the source of constitutive model inadequacy is incorporated in the statistical assessment of model calibration. In comparison with non-intrusive algorithms the classical black-box forward problem is unfolded to identify existing model inadequacies in a physical manner.

The performance of the employed methodology is demonstrated on problems in nonlinear elastography where the identification of the mechanical properties of biological materials can improve non-invasive, medical diagnosis. The discovery of multiple modes and thus inference-solutions as well as quantifying model inadequacy in such problems is crucial for the task of achieving the diagnostic objectives. Finally, impor-

tance sampling is employed in order to verify the results and assess the quality of the provided approximations. It confirms that the bias that is introduced by our method is small. Thus, the overall introduced Bayesian framework in this thesis allows the quantification of uncertainties in high-dimensional large-scale inverse problems.

Zusammenfassung

Die akurate Quantifizierung von Unsicherheiten von modellbasierten, hochdimensionalen inversen Problemen stellt eine der grundlegendsten Herausforderungen der Informatik und Ingenieurswissenschaften dar. Um für eine große Anzahl von Unbekannten statistische Rückschlüsse ableiten zu können sind selbst moderne Stichprobenverfahren nicht anwendbar, da sie eine sehr große Menge an aufwändigen Vorwärtsauswertungen des Modells benötigen. Um dieses Problem zu lösen, werden effiziente und genaue Bayesische Methoden in dieser Arbeit entwickelt. Die neu entwickelten Verfahren basieren auf Bayesischen Variationsmethoden, welche die exakte A-posteriori-Wahrscheinlichkeitsverteilung über die Lösung eines Optimierungsproblems mit Hilfe von passend ausgewählten Wahrscheinlichkeitsverteilungen annähert. Die Inferenz approximiert, indem ein deterministisches Optimierungsproblem gelöst wird. Dies ermöglicht die Berechnung einer guten Abschätzung der exakten A-posteriori- Wahrscheinlichkeitsverteilung mit sehr wenigen Vorwärtsauswertungen.

Die wichtigsten Ziele dieser Arbeit sind es die Limitierungen der gegenwärtigen Algorithmen, die mit dem Fluch der Dimensionalität zurechtkommen müssen, aufzuheben, um Unsicherheiten akkurat zu quantifizieren zu können. Zu Beginn wird ein neues und effektives Verfahren zur Dimensionalitätsreduktion vorgestellt. Dieses Verfahren ermittelt niedrigdimensionale Teilräume, basierend auf einer vollständig Bayesischen Formulierung, in denen die Varianz der angenäherten A-posteriori-Dichte konzentriert ist. Des Weiteren wird die Forderung nach hoher Genauigkeit der A-posteriori Abschätzung adressiert, welche für multimodale A-posteriori-Dichten besonders schwierig zu erfüllen ist. Es wird eine neuartige Bayesische Variationsmethode vorgestellt, die die A-posteriori-Dichte durch eine Kombination von multivariaten Gaußverteilungen annähert. Jede dieser Mischkomponenten liefert eine akkurate, lokale Annäherung der A-posteriori-Dichte, indem es einen individuellen niedrigdimensionierten Teilraum abbildet. Zuletzt wird in dieser Dissertation, neben der Präsenz von Messungenauigkeiten und parametrischer Unsicherheit, eine weitere Quelle der Unsicherheit quantifiziert, die bisher kaum untersucht wurde. Genauer gesagt wird die Quelle von Modellunzulänglichkeiten in die statistische Bewertung der Modellkalibrierung integriert. Im Vergleich zu nicht intrusiven Algorithmen wird das klassische Black-Box Vorwärtsproblem entpackt, um bestehende Modellunzulänglichkeiten auf physikalische Art und Weise zu identi-

fizieren.

Die Leistungsfähigkeit der angewandten Methodik wird anhand von Problemstellungen aus dem Bereich der nichtlinearen Elastografie demonstriert, in der die Identifizierung von mechanischen Eigenschaften von biologischen Materialien die nichtinvasive, medizinische Diagnose verbessern kann. Um die Diagnoseziele zu erreichen, ist es entscheidend, einerseits mehrere mögliche Lösungen zu finden und anderseits die Modellunzulänglichkeiten zu quantifizieren. Abschließend werden mit Importance Sampling die Ergebnisse verifiziert und die Qualität der vorliegenden Abschätzungen überprüft. Es wird bestätigt, dass die Verzerrungen (Bias), die durch unseren Algorithmus eingeführt werden, klein sind. Demnach ermöglichen die in der Dissertation eingeführten Bayesischen Verfahren die Quantifizierung von Unsicherheiten hochdimensionaler inverser Probleme.

Acknowledgements

Over the last few years, many people have helped me to grow, both academically and professionally. First and foremost, I want to thank my research adviser and mentor Phaedon-Stelios Koutsourelakis for his extraordinary support with my research. I am fortunate to have been part of his research group and I would like to thank him for the opportunity of researching in the fascinating field of uncertainty quantification. I am especially grateful for the great opportunity to learn from him, the immense amount of time he has dedicated to my research and for all the faith he has put in me.

Furthermore, I want to thank the members of my committee: Nicholas Zabaras, for the great discussions we have had at conferences, and during your research visits. Thank you for being part the committee and for coming the long way from the US. I also want to thank Wolfgang A. Wall, who has supported me since I was an undergraduate student and who introduced me to the field of computational mechanics. He gave me the opportunity of getting to know his colleague Tarek Zohdi, who first introduced me to the subject of statistical inference in engineering questions. Back then I did not realize how important this topic would become for me.

A special thank you goes to Nassir Navab for the helpful discussions and the possibility to interact with his group members. I also want to thank Sailesh Conjeti and Christoph Hennersperger for testing phantoms, registering images and sharing insights with me.

My thanks also go to all members of the 'Professur für Kontinuumsmechanik'. In particular, I would like to thank my colleagues Michael Kraus and Constantin Grigo for sharing an office with me and for the scientific interchange. A special thanks also goes to Markus Schöberl, who started as the first master student within our group but quickly became a friend and colleague. I would also like to thank my students Lukas Bruder, Christian Heynitz, Mariella Kast, Lukas Köstler, Christoph Moosbauer and Nicola Zimmermann with whom I had great discussions and interesting projects. I would also like to thank all my other friends and colleagues from our faculty for great activities and interactions.

Finally, many thanks are due to my partner, family and close friends for their support throughout this process and over all the years. Thank you very much for being here. You are the best.

Contents

List of Symbols

Mathematical

$\boldsymbol{a}$	boldface, small letter signifies a vector
$\boldsymbol{A}$	boldface, large letter signifies a matrix
$\log(\boldsymbol{A})$	natural logarithm of $\boldsymbol{A}$
$tr(\boldsymbol{A})$	trace of $\boldsymbol{A}$
$\|\boldsymbol{A}\|$	determinant of $\boldsymbol{A}$
$\|\|\boldsymbol{A}\|\|$	Euclidean norm of $\boldsymbol{A}$
$\boldsymbol{A} : \boldsymbol{B}$	scalar product/double contraction of two second order tensors
$\bigtriangledown \boldsymbol{a}$	gradient of $\boldsymbol{a}$ with respect to $\boldsymbol{X}$
$\bigtriangledown \cdot \boldsymbol{a}$	divergence of $\boldsymbol{a}$ with respect to $\boldsymbol{X}$
$\frac{\partial}{\partial \boldsymbol{a}}$	partial differentiation with respect to $\boldsymbol{a}$
a_j	entry j of $\boldsymbol{a}$
$\boldsymbol{a}^{(j)}$	iteration j of $\boldsymbol{a}$

Some deterministic parameters

d_L	number of neighboring pairs of elements	42
d_{FE}	number of finite elements	29
d_f	dimension of forces	112
d_y	dimension of measurements and model output	16
$d_{y,all}$	dimension of displacement, incl. Dirichlet boundary disp.	113
d_Θ	dimension of reduced variables	36
d_Ψ	dimension of model parameters	16
d_S	dimension of discretized stresses or strains	112
Γ_u	Dirichlet boundary	28
Γ_σ	Neumann boundary	28
Ω	physical domain(spatial configuration)	26
Ω_0	physical domain (reference configuration)	26

$\mathbf{x}, \mathbf{X}$	position in the deformed, reference configuration 26
$\boldsymbol{\epsilon}$	vector of strains 113
$\mathbf{n}, \mathbf{N}$	surface normal vector in deformed, reference configuration 27
$\boldsymbol{B}$	linear gradient operator for strains 113
$\hat{\boldsymbol{B}}$	linear gradient operator for distr. stresses 112
$\boldsymbol{\phi}$	mapping from reference to deformed configuration 26
$\mathbf{F}$	deformation gradient 27
$\mathbf{C}$	right Cauchy-Green tensor 27
$\mathbf{D}_e$	local constitutive matrix 113
$\mathbf{E}$	Green-Lagrange strain tensor 27
$\tilde{\boldsymbol{\sigma}}$	Cauchy stress tensor 27
$\mathbf{S}$	second Piola-Kirchoff stress tensor 27
$\boldsymbol{u}$	displacements 26
$\boldsymbol{u}_b$	Dirichlet boundary displacements 113
$\boldsymbol{y}$	model output 34
$\hat{\boldsymbol{y}}$	measurements, observables 16
w	strain energy function 28
J	Jacobian determinant 27
$\boldsymbol{W}$	directions of the lower-dimensional subspace 36
$\boldsymbol{\mu}$	mean value of the representation of $\boldsymbol{\Psi}$ 36
$\boldsymbol{G}$	gradient of the map at $\boldsymbol{\mu}$: $\mathbf{G} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}\|_{\boldsymbol{\Psi}=\boldsymbol{\mu}}$ 40
$\boldsymbol{T}$	parametrization of $\boldsymbol{\mu}, \boldsymbol{W}$: $\boldsymbol{T} = \{\boldsymbol{\mu}_j, \boldsymbol{W}_j\}_{j=1}^{S}$ 72

Probability calculus

$p()$	probability density function 16
$\pi()$	desired probability density function 20
$p_\Psi(\boldsymbol{\Psi})$	prior probability distribution of $\boldsymbol{\Psi}$ 17
$< \boldsymbol{\Psi} >$	expectation of $\boldsymbol{\Psi}$ with respect to $p(\boldsymbol{\Psi})$ 16
$< \boldsymbol{\Psi} >_q$	expectation of $\boldsymbol{\Psi}$ with respect to $q(.)$ 38
$KL(q\|\|p)$	Kullback-Leibler divergence between q and p 16
$\mathcal{N}(\boldsymbol{\Psi}\|\boldsymbol{x}, \boldsymbol{\Sigma})$	Multivariate normal distribution over $\boldsymbol{\Psi}$ with mean $\boldsymbol{x}$ and covariance $\boldsymbol{\Sigma}$ 35
M	number of samples 20
ESS	effective sample size 20
$\mathcal{F}$	lower bound 24

$\boldsymbol{\Psi}$	latent variable (in application: constitutive mat. parameter) 16
$\boldsymbol{\Theta}$	reduced latent variable 36
$\boldsymbol{\eta}$	captures residual variance of dimensionality reduction 37
τ	precision of measurement errors 35
s	discrete latent variable; resp. weight of mix. components . 68
$\boldsymbol{\Upsilon}$	parametrization of $\boldsymbol{y}, \boldsymbol{\sigma}, \boldsymbol{\Psi}, \boldsymbol{H}$: $\boldsymbol{\Upsilon} = \{\boldsymbol{y}, \boldsymbol{\sigma}, \boldsymbol{\Psi}, \boldsymbol{H}\}$ 116
$\boldsymbol{\sigma}$	discretized stress vector 112
$\boldsymbol{v}$	precision of model error 115

Chapter 1

Introduction

> *Most people, if you describe a train of events to them, will tell you what the result would be. They can put those events together in their minds, and argue from them that something will come to pass. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk of reasoning backwards.*
>
> Sherlock Homes to Dr. Watson in: *A study in Scarlet* by Sir Arthur Conan Doyle, 1859-1930 [1].

1.1 Motivation and background

Our everyday decisions are based on predictions of future events which come from extrapolating observations using models of our environment. *Models* and derived inferences are either applied unconsciously, e.g., a child learns many interactions on the fly without explicitly formulating a model, or are constructed by purpose and with expertise. For example, in physics many great researchers, pioneers and explorers developed models to express underlying relationships and correlations over centuries. The models - which are approximate models, based on excessive complexity or partial understanding of the reality - are usually designed to represent the major physical relationships and characteristics of the reality. However, even if we could build exact and perfect models, accurate input parameters are still needed to predict a correct outcome. The connection between the input and output parameters, in the absence of measurement noise,

defines the forward operator. Determining the output with given input parameters relates to solving the *forward or direct* problem.

Inversely, inferring unknown input parameters from probably noisy observations corresponds to an *inverse* problem. Model-based inverse problems appear in many scientific fields, and typically where indirect observations of the quantity of interest are made. A classic example is the mapping from observed arrival times of seismic waves to the earth's subsurface [2, 3]. Other applications are deconvolution problems in astrophysics [4], finding cracks and interfaces in materials [5], the identification of materials within medical tomography [6] or permeability estimation for soil transport processes that can assist in the detection of contaminants, oil exploration and carbon sequestration [7, 8, 9]. In all cases, experimental or computationally-generated data is used, e.g., for model calibration in order to adjust model parameters and to obtain improved predictions. The identification of model parameters can provide insight into the process of interest and feeds the understanding of the system's behavior. Especially in the last one or two decades the field of solving inverse problems increased rapidly, supported by the large increase of computing power and the development of advanced numerical methods [10].

Elastography as an application

This thesis is particularly concerned with the inverse problem of identifying mechanical properties of biological materials based on MRI or ultrasound images in the context of non-invasive medical diagnosis (elastography). The identification of stiffness, or mechanical properties in general, can potentially lead to earlier and more accurate diagnosis, e.g., for breast lesions as malignant or benign tumors [11, 12]. It provides valuable insights to differentiate between modalities of the same pathology [13] and to monitor the progress of treatments.

The term *elastography* refers to techniques which relate medical imaging to elastic properties of soft tissues. It originates from manual palpation, one of the oldest diagnostic methods, which was first evidenced in $400BC$ [14]. Manual palpation assesses the stiffness of a patient's tissue by feeling. A stiff lump, explored by touch, might be an indication for a diseased tissue. Elastography replicates this process. For example, under ultrasound elastography the body is slightly deformed and the internal deformation of the body is tracked by multiple images in order to produce certifiable estimates of the mechanical properties. Testing and comparing the mechanical properties of histologically documented tissues have let to the conclusion that many diseased tissues, such as breast tumors, are of a different stiffness than their surrounding material. The diagnostic value of elastography is based on this opportunity to differentiate tissues based on non-identical mechanical properties. For instance, specific lesions can be identified as fibrous tissue as they are visibly stiffer than normal tissue. Additionally, malignant tissues (ductal carcinoma in situ and invasive ductal carcinoma) can be rec-

ognized and distinguished from benign lesions by their nonlinear stiffness behavior over different strain levels [15, 16]. In Figure 1.1, the different ratios of the elastic moduli for different levels of compression depict that the nonlinearity in the stress-strain relationship varies for dissimilar tissues. In addition, this figure also shows that the magnitude of the stiffness itself varies for different breast tissues.

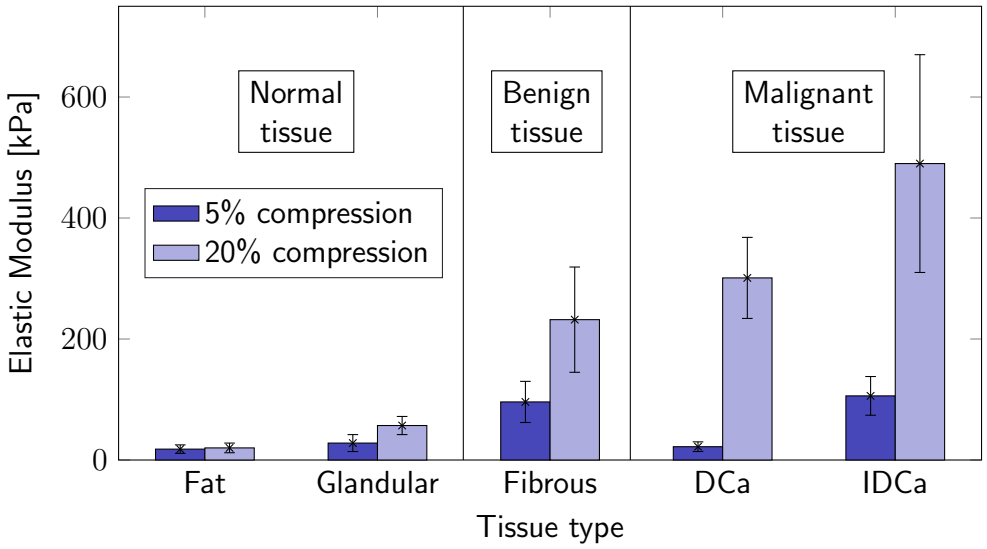


Figure 1.1: Elastic modulus for different breast tissues for different strain/compression levels (in vitro but directly after removal from the body). The elastic modulus magnitude and its variability for different levels of compression can be used to identify the tissue type. DCa is used as an abbreviation of ductal carcinoma in situ and IDCa stands for invasive and infiltrating ductal carcinoma (figure is redrawn based on [16]).

There is a mounting body of evidence that indicates the potential of elastography-based techniques not only for identifying tumors, e.g., breast or prostate tumors [16], but also for other purposes. Among them are, for example, the identification of other diseases, such as the characterization of blood clots [17], atherosclerosis [18], osteopenia [19] and liver fibrosis/cirrhosis [20]. To be more specific, for instance, liver fibrosis increases the stiffness as a whole which is difficult to detect with conventional ultrasound. However, elastography is advantageous as it is able to identify the increase of stiffness. The potential of elastography is especially visible considering the percentage of global causes of death for women. In Figure 1.2, it is shown that 4.94% of causes of death relates to breast cancer, cirrhosis of the liver and liver cancer, which indirectly shows the potential for impact of this method if the diseases would have been treated faster based on an earlier diagnosis. Early detection was identified as a critical factor in increasing survival rates, e.g., in case of breast cancers: the 5-year survival rate for women with stage 0-cancer is 98.8% and with stage 4-cancer around 26.3% [22].

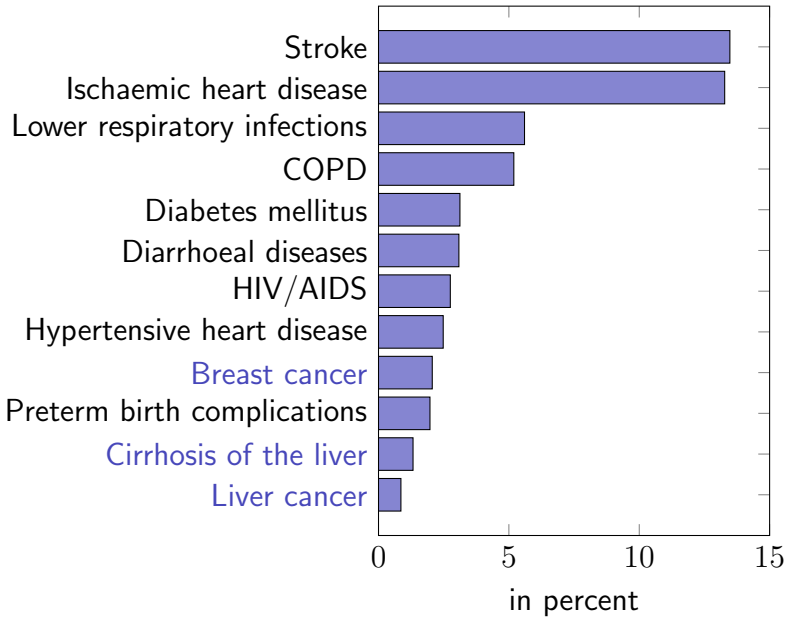


Figure 1.2: Percentage of global causes of death for females in 2012 [21]. COPD is the abbreviation for chronic obstructive pulmonary disease.

Mammograms and conventional ultrasonography represent the primary detection technique to pick up breast lesions. However, both have some restrictions. Mammography performed in dense breasts may often yield false-negative results and ultrasonography is sensitive in the detection of lesions, but its specificity is poor [23]. Many biopsies are performed in benign lesions causing discomfort to the patients and increased costs.

The pioneering work of Ophir and coworkers [24], followed by several clinical studies [25, 26, 27, 28], have demonstrated that the resulting strain images typically improve the diagnostic accuracy over ultrasound alone. Based on its diagnostic potential and increased computational capabilities, it becomes more and more important to develop tools that leverage the capabilities of physics-based models in order to quickly and accurately produce diagnostic estimates with quantified confidence levels.

Elastography is based on an imaging process, e.g., ultrasound, which relates the mechanical properties such as relative stiffness of an unknown tissue to applied forces. This relies on the principle that stiff materials deform less than soft tissues and it is used to inversely derive the mechanical properties from their deformation. All elastographic techniques build upon the following three basic steps [29]: **1)** Exciting the tissue using a (quasi-)static, harmonic or transient source, **2)** Indirectly measuring interior tissue deformation, e.g., displacements, velocities, using an imaging technique such as ultrasound [24], magnetic resonance [30] or optical tomography [31], and **3)** Inferring the mechanical properties from this data by using a suitable continuum mechanical model of the tissue's deformation.

Different elastography techniques can still be grouped by various criteria, as compared in Table 1.1. One classification criterion is based on the measured quantity, which separates strain (quasi-static or dynamic) and shear wave (dynamic) techniques. The *strain-based elastography* technique mechanically deforms the tissue. This is achieved either by an active external displacement of the tissue surface or passively in a physiological way within the tissue. A *strain map* is constructed by calculating the deformation from images at different stages of a compression cycle.
The *acoustic radiation force impulse* (ARFI) technique uses an acoustic radiation force of an ultrasound transducer to perturb the tissue locally at a single location to infer the deformation of the tissue [32, 33].
Within *transient elastography*, an external actuator produces a cycle of low-frequency vibration generating transient shear waves. The velocities of the shear wave are then measured within the tissue. Based on the velocities of the shear waves, the mechanical properties are estimated [34].
In *supersonic shear-wave imaging* (SSI), also called ultra-fast shear wave elastography, the focus of the radiation force in depth changes faster than the speed of the provoked shear waves. An ultra-high frame rate is necessary to track the propagation of the shear waves [35, 34].

	measured quantities	applied force	measuring modalities
Strain elastography	strain	mechanical	images
ARFI	strain	radiation force	images
Transient	shear wave speed	mechanical	point measurements
SSI	shear wave speed	radiation force	images

Table 1.1: Different elastography techniques/products.

Strain elastography is especially advantageous, as it suffices to use just readily available standard ultrasound transducers. This results in a lower relative cost and an increased portability. Within strain imaging, there are two approaches for inferring the constitutive material parameters. In the *direct approach*, the equations of equilibrium are interpreted as equations for the material parameters of interest. The inferred interior strains and their derivatives appear as coefficients and the strain-ratio is then used as a surrogate index for stiffness in the absence of a true index of the material parameters [36, 37]. While such an approach provides a computationally efficient strategy, it does not use the raw data, i.e., noisy displacements. Instead of using raw data, transformed versions are applied, i.e., strain fields or strain derivatives, for example, by application of ad hoc filtering and smoothing data. As a result, the informational content of the data is compromised and the quantification of the effect of observation noise is cumbersome. Furthermore, employed smoothing can smear regions with sharply varying properties

and hinder proper identification. In addition, linear elasticity of the material is assumed.

The alternative to direct methods are indirect or *iterative* procedures which admit the formulation of the inverse problem of interest in this thesis. The unknown input parameters refer to the unexplored material parameters which are inferred by minimizing the discrepancy between observed and model-predicted displacements [38, 39, 40, 41]. In this context, the individual measurement entries are not only used to locally deduce unknown material parameters but to solve a global system. The solution of the global system can also be applied to problems where no interior displacements are measured but only the deformation of the boundary [42]. More importantly, directly incorporated *constitutive laws* allow the use of more specific material models, such as hyperelastic material laws for biomaterials [43, 44, 29]. Thus, the model-based elastography methods can especially be advantageous in distinguishing cancerous lesions or they can also be used as an indicator of the histology, as discussed previously. The solving strategies can be categorized by their optimization method in: Hessian based (Newton method), gradient based or gradient free optimization methods. While these approaches utilize the raw data directly, they generally imply a higher computational cost than the forward problem and potential derivatives of the system response with respect to the input parameters have to be computed several times.

Within this thesis, we employ model-based strain elastography and explore different options to obtain the mechanical properties of the unknown tissue by solving an inverse problem based on derived deformation maps. We aim to develop rigorous, new statistical models and efficient computational tools to quantify the material parameters and their uncertainties, introduced below, for a more precise diagnosis.

Bayesian inference

In most applications, as in model-based elastography, there is no explicit expression for the inverse relation that maps output data to input parameters. Thus, the forward problem and potentially its derivatives with respect to the model variables have to be solved/computed for multiple different plausible input parameters to identify the correct configuration. This can be expensive, especially for a complex system. Accordingly, the solution of *model-based inverse problems* represents a fundamental challenge in the context of model calibration and system identification. Another challenge working with inverse problems is that the problem can be ill-posed, i.e., the solution is not unique, it is sensitive to small perturbations in the data or it might be impossible to perfectly match the outcome of the forward problem with the observations [45]. In addition, significant uncertainties, such as observation noise, usually exist [46]. Another source of uncertainty relates to the incorporation of computational models. Models usually contain simplifications and approximations of the reality and hence inherently include model errors.

The sources of errors and *uncertainties* are schematically depicted in Figure 1.3

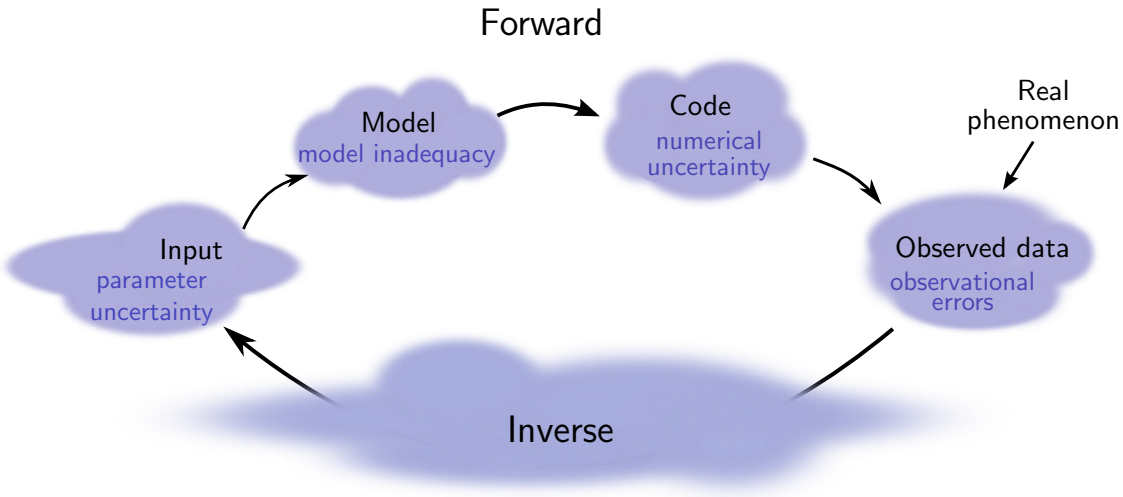


Figure 1.3: Solving the forward or inverse problem is encumbered by different sources of uncertainties.

where cloudy regions represent possible uncertainties.
Firstly, when solving a forward problem to predict an outcome one may be unsure about the correct input parameters. This insufficient knowledge should be included into the algorithm when proposing input parameters. By allowing the model parameters to vary *parameter uncertainty* is incorporated.
Secondly, even if we knew the exact input parameters, the model itself is inaccurate, as no model perfectly reflects the truth. For the majority of applications, the physical relationships are too complex to describe or processes are not understood well enough to design the exact model. Thus, approximate models are designed to represent the major characteristics of the reality. The discrepancy between the outcome of the true and the assumed model relates to *model inadequacy*. Since the process itself may exhibit some natural variations, model inadequacy can be defined as the difference between the true mean value of the real world process and the model output for true input parameters [46].
Thirdly, as the models are usually computationally simulated, *code error* may need to be included. This may relate to numerical fluctuations in the simulation [47]. In addition, even if the numerical code is perfectly written and free of errors, it may be impractical to run the code at all possible combinations of input parameters. Thus, interpolation errors based on interpolated model output may occur, which should also be taken into the consideration [46].
Finally, when we want to predict observations we have to also incorporate possible *observation errors*, which outline the difference between the actual observation and the

true outcome.
Therefore, in order to correctly infer model parameters it is essential to incorporate all the existing uncertainties, when solving inverse problems.

Solving the inverse problem and quantifying uncertainties is a subject of the research field of *Bayesian inference*. Bayesian formulations offer a rigorous setting for their solution as they account for various sources of uncertainty that is unavoidably present in these problems. Furthermore, they possess a great advantage over deterministic alternatives: Apart from point-estimates they provide quantitative metrics of the uncertainty in the unknowns, encapsulated in the *posterior distribution* [48].

Nonetheless, the solution of model calibration problems in the Bayesian framework is hampered by multiple difficulties:
Firstly, for high-dimensional problems an exuberant number of computationally expensive forward calls are required, which poses a prohibitive computational burden.
Secondly, multimodal probability distributions offer many local maxima, which is difficult for algorithms to correctly identify. For example, Markov chains get caught for extended periods of time in local maxima or local approximations schemes are only able to record a single mode.
Thirdly, an obstacle which is, despite its importance, usually ignored is model inadequacy. It is often assumed that the model, for example used for calibration, is perfect. This leads to misidentified model parameters or, even worse, wrong predictions. This thesis addresses these major challenges and proposes a novel, efficient and accurate framework. Before deriving details about the main work contributions, a short review of related work is outlined.

1.2 Related work

Computational efficiency and dimensionality: Solving large-scale inverse problems is computationally very expensive, if not an intractable process. Finding a solution for an inverse problem with standard Markov Chain Monte Carlo (MCMC, [49]) techniques requires an exorbitant number of likelihood evaluations in order to converge, i.e., solutions of the forward model [50, 51, 52, 53], Section 2.2.3. The large number of required forward calls originates from the poor scaling of traditional Bayesian inference tools with respect to the dimensionality of the unknown parameter vector - another instance of the *curse-of-dimensionality* [54]. As each of these calls implies the solution of very large systems of (non)linear equations, those approaches are usually impractical for high-dimensional problems. In problems, such as the elastography example, the model parameters of interest, i.e., material properties, exhibit spatial variability which requires fine discretizations in order to be captured. Consequently, the solution of large-scale inverse problems critically depends on methods to reduce computational cost. Several authors, such as T. Bui-Thanh, T. Cui, O. Ghattas, N. Petra, Y.M.

Marzouk, G. Stadler, L.C. Wilcox and many more, work on different efficient methods, either achieved by reducing the number and/or the cost of a single required forward call, which will be briefly summarized.

Advanced sampling schemes, like adaptive MCMC [55, 56, 57] and Sequential Monte Carlo (SMC, [58, 59, 60]) exploit the physical insight and the use of multi-fidelity solvers in order to expedite the inference process. The use of first-order derivatives, Hessian [61] or low-rank structure of the Hessian [62, 63] to design effective proposal distributions has also been advocated either in a standard MCMC format or by developing advanced sampling strategies [64]. These are generally available by solving appropriate *adjoint problems* which are well-understood in the context of deterministic formulations. Nevertheless, the number of forward calls can still be in the order of tens of thousands if not even higher.

Several propositions have also been directed towards using emulators, surrogates or reduced-order models of various kinds [65, 66, 67, 68, 69, 70, 71, 72, 73]. The forward model is replaced with an inexpensive surrogate to dramatically decrease the computational cost of a forward call. However, such a task is severely hindered by the high-dimensionality. More recent methods attempt to exploit the lower-dimensional structure of the target posterior where maximal sensitivity is observed [74, 75, 76, 73, 77, 78]. This enables inference tasks carried out on spaces of significantly reduced dimension and are not hampered by the aforementioned difficulties. Generally, all such schemes construct approximations around the maximum a posteriori (MAP) point by employing local information, e.g., based on gradients, and are therefore not suitable for multimodal or highly non-Gaussian posteriors.

An alternative to sampling approaches are non-empirical approximation schemes, such as Variational Bayesian (VB) [79, 54], see Section 2.2.4, which reduces the number of expensive forward calls. Such methods have risen into prominence for probabilistic inference tasks in the machine learning community [80, 81, 82] but have recently also been employed in the context of inverse problems [83, 84]. They provide *approximate* inference results by solving an optimization problem over a family of appropriately selected probability densities with the objective of minimizing the Kullback-Leibler divergence [85] with the exact posterior. The success of such an approach hinges upon the selection of appropriate densities that have the capacity of providing good approximations while enabling efficient and preferably closed-form optimization with respect to their parameters. Based on the great advantages of Variational Bayesian frameworks advanced VB strategies are employed in this thesis resolving the remaining challenges. We note that an alternative optimization strategy, originating from a different perspective and founded on map-based representations of the posterior, has been proposed in [86].

Multimodality: *Multimodal* posteriors are not only a challenge for local approximation schemes but also for standard MCMC methods. Multimodality often causes

mixing problems as the Markov chain is trapped in minor modal areas for long periods of time. This is especially exacerbated for higher dimensions. Different advanced inference tools [87, 88, 89, 90], such as those based on simulated annealing, annealed importance sampling or nested sampling, have been developed. However, they require a very large number of forward model calls, increasing with the number of unknowns.

Alternatively, different mixture models have been developed in various statistical inference applications, e.g., speaker identification [91], data clustering [92], and also in combination with Variational Bayesian inference techniques [93, 79, 94]. Nevertheless, all of these problems are characterized by inexpensive likelihoods, low-dimensional problems and multiple data/measurements. In this thesis, a model is developed that overcomes existing problems with a mixture of Gaussians within an advanced novel Variational Bayesian framework. It is able to solve computationally expensive, high-dimensional problems with a physical collection of data in a single test/experiment.

Model inadequacy: Another challenge, which has so far barely been accounted for, is *model inadequacy*. In most model calibration studies, it is implicitly assumed that the model is perfect. However, physical systems are very complex and simple mathematical models are used to approximate the reality. The Bayesian framework allows the comparison of different models for model selection, e.g., with Bayes factor [95] or information criteria [96, 97]. Nonetheless, none of these methods explicitly quantify the model error, nor do they provide a predictive uncertainty that is representative of the extent of the model error. They merely compare different models with each other. Different approaches [46, 98, 99, 47] to quantify the model error explicitly model the model error as an additive term to the model outcome, e.g., by a Gaussian process. In the view of the fact that he discrepancy model is posed only on the observables quantities it is fine-tuned with respect to these observations. Thus, it does not provide much physical insights on model error and does not significantly improve the predictive capabilities of the model [34]. In addition, it gets entangled with the measurement errors and a disambiguation of model and data error is difficult. Moreover, multiple physical experiments are required and it becomes problematic for high-dimensional problems.

In contrast to that, Berliner [100] was one of the very first who embedded an additive term within a submodel to quantify model error for an example with a small number of unknown model parameters. This approach of embedding the model error in a submodel has been extended in the field of fluid dynamics for large-scale problems. More specifically, within approximate turbulence models, such as RANS, Boussinesq approximation, Spalart-Allmaras (SA) or the $k - \omega$ turbulence model an additional term, e.g., Gaussian process, is added within the approximate model [101, 102, 103, 104, 105, 106]. The model error for a specific problem is then identified by comparing the outcome of the approximate model with the outcome derived by direct numerical simulation (DNS). It relies on the assumption that the outcome of the DNS, which is computationally very expensive to derive, is the 'true' outcome [107, 104]. In those

cases, only the model error is treated as an unknown and no further latent or model parameters are quantified. This transfers the problem to a model calibration problem where the model error can be interpreted as a model parameter.

In this thesis, a new developed strategy of identifying model inadequacy is presented. This strategy is based on a framework by Koutsourelakis [108] but extends it with a consistent derivation of the normalization term, such that the integration of flexible prior assumptions is possible. The intrusive framework unravels the forward problem, which enables us to assess constitutive model error directly without knowing the true model. In contrast to existing work, the quantification of the model parameters and the model error is at the same time possible making inference for high-dimensional problems feasible.

1.3 Work contributions and broader impact

This thesis focuses on computational methods for large-scale nonlinear inverse problems in a Bayesian framework and addresses the previously explored main three challenges: Computational efficiency and 'curse of dimensionality', assessing multimodality and model inadequacy. The main contributions are threefold:

Firstly, we investigate a computationally efficient Variational Bayesian framework, directed towards approximating the exact posterior by solving a deterministic optimization problem. Specifically, we propose a dimensionality reduction of the unknown parameters capturing as much as possible of the associated posterior density. We elaborate on the lower-dimensional structure of the target posterior by identifying subspaces where most of the probability mass is contained. This is achieved by using a fully Bayesian argumentation resulting in a highly efficient framework which enables the solution of high-dimensional nonlinear inverse problems.

Secondly, we propose a Variational Bayesian strategy to capture multimodal probability distributions. In contrast to existing approaches [93, 79, 94] the inverse problems considered here are based on a single experiment [109]. Subsequently, we use mixtures of Gaussians to approximate the posterior for model-based high-dimensional inverse problems.

Thirdly, we intrusively quantify constitutive model inadequacy in a large-scale inverse problem. In contrast to non-intrusive state-of-the-art algorithms we open the classical black-box forward problem and bring all model equations to the forefront to identify existing model inadequacies in a physical manner. As a result, the constitutive model error can be locally quantified. Note that physical constraints are satisfied at the same time. This direct estimate of the model inadequacy can then be used for predictive estimates. All this is approached by employing a fully Bayesian formulation.

In this thesis, the developed framework is demonstrated in nonlinear elastography. Up until now, uncertainties within this application have barely been considered. Just

recently, uncertainties in the image registration process have been examined [110, 111, 112, 113] whereas the group of Risholm [114] incorporates uncertainty quantification also for lung elasticity estimation. The authors [108, 115] extended the approaches of quantifying unknown tissue parameters and their uncertainties, e.g., by incorporating a dimensionality reduction with radial basis functions. However, in all cases linear elastic materials are assumed and computationally expensive sampling schemes are employed. In contrast to that, in this thesis a fast and computationally cheap, but accurate quantification of the unknown material parameters and the uncertainties for high- dimensional nonlinear problems is proposed. It provides an accurate diagnosis within the application: elastography. This methodology can be used to 1) reduce the required number of performed biopsies in benign lesions which causes discomfort to patients and increases costs and 2) minimize the number of false-negative results.

These developments are significant because they also contribute to the foundations of interdisciplinary science of Bayesian inference for large-scale inverse problems, e.g., for problems from engineering and medical sciences. Only software interfaces with the outcome of the forward call and its derivatives need to be available. Then, the new developed framework, incorporating dimensionality reduction, capturing multimodal posteriors accurately and quantifying constitutive model error, can directly be applied. For a different constitutive model the framework of quantifying model inadequacy needs further research. Specific physical insight can be straightforwardly integrated by priors.

1.4 Outline of the thesis

Chapter 2 contains the fundamental core of the thesis. It introduces the required basics of uncertainty quantification with a focus on different approximation methods. In addition, a short overview of solid and computational mechanics is given, which is used to build the forward model for elastography.

Chapter 3 investigates a novel framework of Variational Bayes for the solution of nonlinear inverse problems incorporating a dimensionality reduction technique. The new developed framework is able to compute the posterior with very few forward calls and is able to find a lower-dimensional subspace where a good approximation of the posterior can be obtained. This can be achieved with a fully Bayesian argumentation. Information-theoretical criteria are developed to identify the cardinality of the reduced coordinates. The performance of the framework is demonstrated for problems of nonlinear elastography. However, the presented methods can also be applied to various other applications. For verification purposes, importance sampling is employed and shows the efficacy of the provided approximation.

Chapter 4 is an extension of the previous chapter with the ability to capture and identify multimodal posteriors. The proposed Variational Bayesian-based strategy approximates the posterior with a mixture of multivariate Gaussians. For each Gaussian,

a lower dimensional subspace is identified, where the posterior is mostly concentrated. The framework is applied to static, nonlinear elastography where a multimodal approximation provides a more accurate picture to the analyst such that better diagnostic decisions can be drawn. Lastly, importance sampling is involved for verification, showing that the introduced bias by the approximation is small and can efficiently be corrected.

Chapter 5 proposes a new strategy of identifying model inadequacy with an unfolded 'black-box' approach. The intrusive framework unravels the forward problem which enables us to assess constitutive model error directly. Again, a Variational Bayesian formulation is included for computational efficiency. Specific problems are analyzed and discussed for the application to elastography.

Chapter 6 concludes this thesis with a summary of the main contributions and investigated ideas. In addition, some future research recommendations are discussed.

Chapter 2

Fundamentals

> *An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*

John W. Tukey, 1915-2000 [116].

Our research objective is to assess unknown parameters based on observations (inverse problem) and to quantify underlying uncertainties for high-dimensional problems. Although the novel framework is generally applicable, in this work it is chiefly applied to the application elastography (Chapter 1). In this application, unknown material parameters are inferred from observed deformation maps, e.g., by solving the inverse problem. In this thesis, the forward continuum mechanics model considers nonlinear elasticity material models and large deformations. Since in our application different disciplines, such as uncertainty quantification and computational mechanics are combined, the aim of this chapter is to summarize required fundamental pillars within each discipline and to provide further references. Within the first subsection, the basic rules of probability theory are reviewed before addressing the variety of methods for employing approximate inference when an analytic solution is not achievable. Furthermore, advantages and disadvantages of the different inference schemes are described. In the end of the first subsection, we compare *Variational Bayes* – which is an approximate inference method – to state-of-the-art inference schemes. In the second subsection, the basics of computational mechanics, especially of nonlinear solid mechanics, are summarized. Those methods will be incorporated later on within our application of interest, elastography.

2.1 Basic concepts of probability theory

Probability theory concerns with probability and random events in the field of mathematics. Starting from the three axioms of probability, it covers discrete and continuous probability distributions and how expectation values of random variables are related to them. It also includes various strategies in cases where such expectation values cannot be derived analytically. We will elaborate on the last one, approximate inference, in more detail. For completeness, we briefly cover a few basics and introduce notational conventions. We refer readers for more introductory literature to [117, 118, 54] and for more advanced relevant work to [54, 119].

Within this thesis, all random variables are continuous (except one: s, see Section 4.2.1). Therefore, we focus on the characteristics of continuous random variables in the following. Let $p(.)$ denote a continuous probability density function, $\mathbf{\Psi}$ a vector of random variables and $g(\mathbf{\Psi})$ a function of $\mathbf{\Psi}$. It is often of interest to describe a probability distribution using, e.g., its first and second moments. In a general way, they can be derived by taking the expectation

$$< g(\mathbf{\Psi}) >=< g(\mathbf{\Psi}) >_{p(\mathbf{\Psi})}= \int g(\mathbf{\Psi})\ p(\mathbf{\Psi})\ d\mathbf{\Psi}, \tag{2.1}$$

with $g(\mathbf{\Psi}) = \mathbf{\Psi}$ for the first moment/mean value and $g(\mathbf{\Psi}) = (\mathbf{\Psi} - < \mathbf{\Psi} >)^2$ for the second moment/variance. For notational economy, the index $_{p(\mathbf{\Psi})}$ is usually omitted unless the expectation is derived under a different probability distribution.

Another important quantity within this thesis is the *Kullback-Leibler* (KL) divergence $KL(q(\mathbf{\Psi})||p(\mathbf{\Psi}))$ between two probability distributions $q(\mathbf{\Psi})$ and $p(\mathbf{\Psi})$. Emerging from the field of information theory it is also called relative entropy. The KL-divergence can be used as a measure of the difference between the two probability distributions $q(\mathbf{\Psi})$ and $p(\mathbf{\Psi})$ [120]:

$$KL(q(\mathbf{\Psi})||p(\mathbf{\Psi})) = -\int q(\mathbf{\Psi}) \log \frac{p(\mathbf{\Psi})}{q(\mathbf{\Psi})}\ d\mathbf{\Psi} = - < \log \frac{p(\mathbf{\Psi})}{q(\mathbf{\Psi})} >_{q(\mathbf{\Psi})} . \tag{2.2}$$

By definition, the KL-divergence is non-negative, becomes zero if and only if $q(\mathbf{\Psi}) = p(\mathbf{\Psi})$ and is a nonsymmetric quantity: $KL(q(\mathbf{\Psi})||p(\mathbf{\Psi})) \neq KL(p(\mathbf{\Psi})||q(\mathbf{\Psi}))$.

So far, we have described the definitions of expectations and the KL-divergence, but we have not mentioned yet any relation of random variables to observed data. Let the observed data be designated by $\hat{\boldsymbol{y}} \in \mathbb{R}^{d_y}$ and the random and unknown parameters of a given model by $\mathbf{\Psi} \in \mathbb{R}^{d_\Psi}$. d_y is the dimension of the measurements and d_Ψ the dimension of the model parameters. The probability of having a measured output $\hat{\boldsymbol{y}}$ based on the parameters $\mathbf{\Psi}$ is expressed by the *likelihood* $p(\hat{\boldsymbol{y}}|\mathbf{\Psi})$. This likelihood includes the model one is investigating within physical applications. The likelihood can be combined with prior beliefs to draw conclusions on unobserved quantities based on

observed data which is expressed by the posterior. The *posterior* $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$ is the resulting conditional probability of $\boldsymbol{\Psi}$ conditioned on the observables $\hat{\boldsymbol{y}}$. Originally formulated in the 18th century by Thomas Bayes [121], and in a general form by Laplace [122], the relation between prior, likelihood and posterior is described by the *Bayes rule*:

$$posterior = \frac{likelihood \times prior}{evidence}, \quad p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}) = \frac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi})\ p_{\Psi}(\boldsymbol{\Psi})}{p(\hat{\boldsymbol{y}})}, \tag{2.3}$$

which follows directly from the product rule ($p(\hat{\boldsymbol{y}}, \boldsymbol{\Psi}) = p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi})\ p(\boldsymbol{\Psi})$). The denominator, the evidence, does not depend on $\boldsymbol{\Psi}$. It is thus often omitted. This yields the unnormalized posterior:

$$p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}) \propto p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi})\ p_{\Psi}(\boldsymbol{\Psi}). \tag{2.4}$$

The evidence is based on the total law of probability $p(\hat{\boldsymbol{y}}) = \int p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi})\ p(\boldsymbol{\Psi})\ d\boldsymbol{\Psi}$. Usually, one would explicitely denote the dependency on a given model M within the prior, likelihood, evidence or posterior distribution. Nonetheless, we omit it for simpler notation. Prior beliefs on $\boldsymbol{\Psi}$ before any $\hat{\boldsymbol{y}}$ are observed are described by a *prior distribution* $p_{\Psi}(\boldsymbol{\Psi})$. This distribution can be subjective. Priors can be separated in *informative* and *non-informative* or in *conjugate* and *non-conjugate* priors. Informative priors convey some specific information, e.g., based on personal experience, insight, historical data. In contrast to this, non-informative priors only possess vague and general information, e.g., a uniform distribution on the normal mean including all possible values of $\boldsymbol{\Psi}$. For details about the advantages and disadvantages of non-informative priors we refer to [48]. A prior is called conjugate to a likelihood if the posterior belongs to the same family as the prior. For instance, a Gaussian prior on an unknown mean of a Gaussian likelihood is conjugate and the posterior is thus Gaussian as well. A mixture of conjugate priors is also conjugate. Conjugate priors are advantageous for computational tractability as posterior distributions become simple.

2.2 Approximate inference

For most problems *exact inference* on the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$ is intractable. This can, for instance, be caused by high dimensional or highly complex posterior distributions. Therefore, many approximation techniques have been developed with different accuracy and computational cost, such as point estimates, local approximations or Monte Carlo methods, see Figure 2.1. For verification and comparison purposes, we will shortly summarize the main aspects of major algorithms (ordered by increasing computational cost) before discussing Variational Bayesian approximation - mainly used within this thesis - in detail.

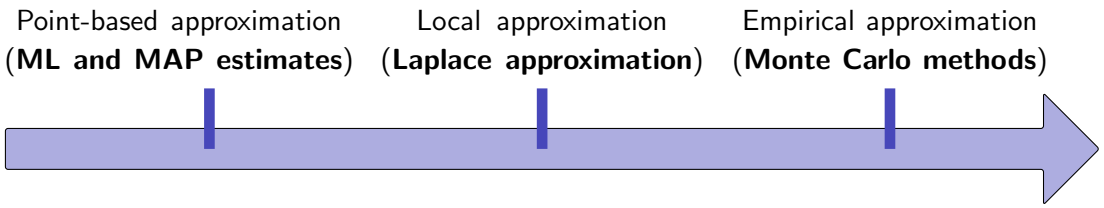


Figure 2.1: Increase of approximation quality and computational cost.

2.2.1 Point-based approximation

Point-based approximations, such as maximum likelihood (ML) and maximum a posteriori (MAP), are one of the most rough approximations to Bayesian inference [123, 79]. Within the maximum likelihood or maximum a posteriori estimation, the value $\boldsymbol{\Psi}$ is identified maximizing the likelihood or posterior. The ML estimate,

$$\boldsymbol{\Psi}_{ML} = \arg \max_{\boldsymbol{\Psi}} p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi}) = \arg \max_{\boldsymbol{\Psi}} \log p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi}), \tag{2.5}$$

can be heavily biased, especially for a small number of samples. The MAP parameters are found by:

$$\begin{aligned} \boldsymbol{\Psi}_{MAP} = \arg \max_{\boldsymbol{\Psi}} p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}) &= \arg \max_{\boldsymbol{\Psi}} \log p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}) \\ &= \arg \max_{\boldsymbol{\Psi}} (\log p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi}) + \log p_{\Psi}(\boldsymbol{\Psi})). \end{aligned} \tag{2.6}$$

There are various drawbacks to point estimates [124]: Most importantly, they provide no measure of uncertainty and are not very representative of the underlying distribution. In addition, using the MAP estimate can result in overfitting and one is very likely to be over-confident of the predictions made by the MAP model. Furthermore, it is problematic that the point estimates are not invariant to reparameterization of the probability distribution.

However, the normalization constant of the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$, which may be difficult to compute, is not required for the computation of $\boldsymbol{\Psi}_{MAP}$, as the normalization constant does not depend on $\boldsymbol{\Psi}$ [125]. The ML or MAP estimates can be calculated in many different ways: By deterministic numerical optimization (e.g., conjugate gradient or Newton's method, which may be sensitive to starting values) or by statistical methods, such as Monte Carlo methods, using simulated annealing. In any case, deriving a ML or MAP estimate is computationally cheap compared to other approximation methods.

2.2.2 Laplace approximation

A simple approximation framework which is also often used is the so called Laplace approximation [126]. It approximates the continuous probability distribution to be esti-

mated - referred to as the posterior in this thesis - by a Gaussian around the maximum (MAP estimate): $q(\boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{\Psi}_{MAP}, \boldsymbol{H}^{-1})$. The analytical, local approximation can be derived by a Taylor series around $\boldsymbol{\Psi}_{MAP}$:

$$\begin{aligned}\log p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}) &= t(\boldsymbol{\Psi}) \\ &= t(\boldsymbol{\Psi}_{MAP}) + (\boldsymbol{\Psi} - \boldsymbol{\Psi}_{MAP})^T \left.\frac{\partial t(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}}\right|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}_{MAP}} \\ &+\tfrac{1}{2}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_{MAP})^T \left.\frac{\partial^2 t(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}\partial \boldsymbol{\Psi}}\right|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}_{MAP}} (\boldsymbol{\Psi} - \boldsymbol{\Psi}_{MAP}) \\ &+... \\ &\approx t(\boldsymbol{\Psi}_{MAP}) + \tfrac{1}{2}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_{MAP})^T \boldsymbol{H}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_{MAP}),\end{aligned} \tag{2.7}$$

with $\boldsymbol{H}$ the Hessian of the log-posterior at $\boldsymbol{\Psi}_{MAP}$. The linear term vanishes since the first order derivatives are zero at $\boldsymbol{\Psi}_{MAP}$. Compared to point-based approximations, such as MAP, the Laplace approximation also estimates the underlying uncertainties. The only additional computational cost is the computation of the Hessian (or the approximation of it) at $\boldsymbol{\Psi}_{MAP}$, since for MAP point estimates the normalization constant of the true distribution is not required [54]. In general, a Gaussian approximation becomes more accurate for an increasing number of experiments (central limit theorem [118, 119]). Nevertheless, the Laplace approximation is unable of approximating multimodal distributions. Furthermore, Gaussian approximations are poorly suited to positive or constrained parameters, e.g., precisions, as it assigns non-zero mass outside the parameter domain. This can be avoided by a reparameterization [127] . Nonetheless, the location of the maximum of $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$ is not invariant to a nonlinear reparameterization. For more information about the Laplace approximation and its characteristics the reader can result [95, 54, 79].

2.2.3 Empirical approximation - Monte Carlo methods

In the previous two subsections, we reviewed inference approximations which are relatively cheap to calculate numerically but can be far off the real solution. Which alternatives do exist if the Gaussian approximation (Laplace approximation) is inadequate and if computationally more expensive methods can be applied? Although for some applications the posterior itself is of interest, mostly integrals (e.g., estimates of the first or second moments, see Equation (2.1)) need to be evaluated. For example, the expectation of $g(\boldsymbol{\Psi})$ with respect to the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$, is expressed by the integral I:

$$I = \int g(\boldsymbol{\Psi})\ p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})\ d\boldsymbol{\Psi}. \tag{2.8}$$

Monte Carlo methods are numerical integration methods which offer a general approach to approximate integrals. They are, for example, employed for inference, model validation or prediction, which is often of primary interest. To approximate the integral

in Equation (2.8) by a Monte Carlo method, a *(pseudo-) random number generator* is required. The *Monte Carlo algorithm* using M number of samples is:

- sample $m = 1 : M$ samples $\mathbf{\Psi}^{(m)}$ from $p(\mathbf{\Psi}|\hat{\boldsymbol{y}})$
- unbiased estimate of the integral is given by

$$I \simeq I_M = \frac{1}{M} \sum_{m=1}^{M} g(\mathbf{\Psi}^{(m)}),$$

which converges by the *Strong Law of Large Numbers* for $M \to \infty$ to: $I_M \to I$. To run Monte Carlo, it suffices to be able to draw samples from $p(\mathbf{\Psi}|\hat{\boldsymbol{y}})$ and to evaluate $g(\mathbf{\Psi})$. An increasing amount of samples increases the accuracy and in the limiting case, $M \to \infty$, one obtains the exact value of the integral. However, Monte Carlo methods can be computationally very expensive as $g(\mathbf{\Psi})$ needs to be evaluated many times.

Later on, we will use some sampling schemes for comparison/verification purposes. For this reason, we briefly discuss the importance of sampling and the general Metropolis algorithm in more detail. Additionally, we derive the (normalized) *effective sample size* (ESS) for both algorithms. The ESS is used to measure the efficiency of the algorithms and provides a measure of comparison with other inference strategies. The ESS determines how informative a given sample is and takes values between $\frac{1}{M}$ and 1 [128]. An $ESS = 1$ relates to an algorithm which is highly efficient compared to one with an $ESS = \frac{1}{M}$ (compare also the ESS_{IS} and ESS_{MCMC} below).

Importance Sampling

In situations when it is not straightforward to sample from a desired probability distribution $\pi(\mathbf{\Psi})$, here $\pi(\mathbf{\Psi}) = p(\mathbf{\Psi}|\hat{\boldsymbol{y}})$, but an evaluation of $\pi(\mathbf{\Psi})$ is easy for a given $\mathbf{\Psi}$, one can use *importance sampling* (IS). In IS, one samples from an auxiliary distribution $q(\mathbf{\Psi})$ and then corrects the estimate by weights, e.g., to approximate the integral:

$$\begin{aligned} I &= \int g(\mathbf{\Psi})\ \pi(\mathbf{\Psi})\ d\mathbf{\Psi} = \int [g(\mathbf{\Psi})\ \tfrac{\pi(\mathbf{\Psi})}{q(\mathbf{\Psi})}]\ q(\mathbf{\Psi})\ d\mathbf{\Psi} \\ &\simeq I_M = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} g(\mathbf{\Psi}^{(m)})\ \tfrac{\pi(\mathbf{\Psi}^{(m)})}{q(\mathbf{\Psi}^{(m)})}. \end{aligned} \tag{2.9}$$

The two steps of *importance sampling* are:

- sample $m = 1 : M$ samples $\mathbf{\Psi}^{(m)}$ from $q(\mathbf{\Psi})$
- approximate the integral by

$$I \simeq I_M = \frac{1}{M} \sum_{m=1}^{M} g(\mathbf{\Psi}^{(m)})\ w^{(m)}, \tag{2.10}$$

where $w^{(m)} = \frac{\pi(\mathbf{\Psi}^{(m)})}{q(\mathbf{\Psi}^{(m)})}$ are the corresponding *importance weights*.

Belonging to Monte Carlo methods, IS converges by the *Strong Law of Large Numbers* for $M \to \infty$ to: $I_M \to I$.

The (normalized) effective sample size for importance sampling can be expressed using the importance weights (details in [129]):

$$ESS_{IS} = \frac{(\sum_{m=1}^{M} w^{(m)})^2}{M \sum_{m=1}^{M} (w^{(m)})^2}. \tag{2.11}$$

It points out the percentage of number of samples that actually contribute to the estimate. The latter attains values between the following extremes: In one extreme (when $ESS \to \frac{1}{M}$) a single sample has a unit normalized weight, whereas the others have zero weights. That happens if $q(\mathbf{\Psi})$ provides a poor approximation and the ESS is dominated by the largest weight $w(\mathbf{\Psi}^{(m)})$. In the other extreme, when $q(\mathbf{\Psi})$ coincides with the exact posterior, all samples have equal weights $w(\mathbf{\Psi}^{(m)})$ and are equally informative ($ESS \to 1$).

The performance of importance sampling can decay rapidly in high dimensions [130, 79]. Therefore, we discuss a very general and powerful algorithm in the next subsection, the Markov Chain Monte Carlo method.

Markov Chain Monte Carlo

Importance sampling can be very inefficient (if the proposal distribution is badly selected) and suffers from severe limitations in high-dimensional problems. Interesting and often used alternatives to importance sampling are *Markov Chain Monte Carlo* (MCMC) methods which combine Markov chains with Monte Carlo techniques to focus on more important regions. Within MCMC, a chain of a correlated (and therefore not independent) sequence of samples is generated starting from any configuration $\mathbf{\Psi}^{(1)}$. Each sample is a non-deterministic function of its previous sample $\mathbf{\Psi}^{(m)} \xrightarrow{T^{(m)}} \mathbf{\Psi}^{(m+1)}$ (only of the previous sample, following the conditional independence property of Markov chains). The construction of samples is based on the usage of a transition kernel $T^{(m)}(\mathbf{\Psi}^{(m)}, \mathbf{\Psi}^{(m+1)}) = p(\mathbf{\Psi}^{(m+1)}|\mathbf{\Psi}^{(m)})$ in such a way that $\mathbf{\Psi}^{(m+1)} \sim T^{(m)}(\mathbf{\Psi}^{(m)}, \mathbf{\Psi}^{(m+1)})$. To ensure convergence of a Markov chain towards the desired probability distribution in the limit of a large number of samples, it needs to be invariant with respect to the Markov chain. A more restrictive condition to ensure invariance is the fulfillment of detailed balance, $\int p(\mathbf{\Psi}^{(m)})\, T^{(m)}(\mathbf{\Psi}^{(m)}, \mathbf{\Psi}^{(m+1)})\, d\mathbf{\Psi}^{(m)} = p(\mathbf{\Psi}^{(m+1)})$ where $p(\mathbf{\Psi})$ is the distribution which the samples have to follow. A Markov chain respecting the detailed balance is also reversible. More information, also about π-irreducibility and aperiodicity, can be found in [54, 128, 119]. Traditionally, $T^{(m)}$ is generated using a proposal probability distribution $\mathbf{\Psi}^* \sim q(\mathbf{\Psi}|\mathbf{\Psi}^{(m)})$ dependent on the previous sample. In cases of a non-symmetric proposal distribution

($q(\mathbf{\Psi}^{(m)}|\mathbf{\Psi}) \neq q(\mathbf{\Psi}|\mathbf{\Psi}^{(m)})$), one needs to incorporate the ratio $\frac{q(\mathbf{\Psi}^{(m)}|\mathbf{\Psi}^*)}{q(\mathbf{\Psi}^*|\mathbf{\Psi}^{(m)})}$ within the calculation of the acceptance ratio $\frac{\pi(\mathbf{\Psi}^*)}{\pi(\mathbf{\Psi}^{(m)})}$to ensure reversibility.

One of the simplest MCMC methods is the *general Metropolis* algorithm which we briefly describe below as we refer to it later. Similar to importance sampling, one samples from a proposal distribution $q(\mathbf{\Psi}|\mathbf{\Psi}^{(m)})$, now depending on the current sample $\mathbf{\Psi}^{(m)}$. In the general Metropolis algorithm, one iterates the following steps M times:

- sample $\mathbf{\Psi}^* \sim q(\mathbf{\Psi}|\mathbf{\Psi}^{(m)})$
- generate a random number $\alpha \sim Uniform[0,1]$ and
  - accept $\mathbf{\Psi}^*$ for $\alpha < \frac{\pi(\mathbf{\Psi}^*)}{\pi(\mathbf{\Psi}^{(m)})}$: $\mathbf{\Psi}^{(m+1)} = \mathbf{\Psi}^*$
  - or otherwise: $\mathbf{\Psi}^{(m+1)} = \mathbf{\Psi}^{(m)}$
- $m \leftarrow m+1$.

The integral in Equation (2.8) can be approximated similarly to Equation (2.9) by

$$I \simeq I_M = \frac{1}{M}\sum_{m=1}^{M} g(\mathbf{\Psi}^{(m)}). \tag{2.12}$$

Within the general Metropolis algorithm, the proposal distribution is also symmetric: $q(\mathbf{\Psi}^{(m)}|\mathbf{\Psi}^{(m+1)}) = q(\mathbf{\Psi}^{(m+1)}|\mathbf{\Psi}^{(m)})$, $\forall \mathbf{\Psi}^{(m)}$.

The convergence rate inversely scales with the square root of the number of parameters [128]. The constant in front of this rate highly depends on how well the proposal step fits the specific problem and other algorithmic details. The autocovariance $\rho(k)$ at lag k can be used to evaluate the independence of the consecutive sample draws [131]:

$$\rho(k) = \frac{1}{(M-k)\rho_0}\sum_{m=1}^{M-k}(\Psi^{(m)} - \bar{\Psi})(\Psi^{(m+k)} - \bar{\Psi}), \tag{2.13}$$

where $\bar{\Psi}$ is denoted as the mean of Ψ and $\rho_0 = \frac{\sum_{m=1}^{M}(\Psi^{(m)}-\bar{\Psi})^2}{M}$. The (integrated) autocorrelation time τ_{int} can be computed as [51, 132, 119]:

$$\tau_{int} = 1 + 2\sum_{k=1}^{\infty}\rho(k). \tag{2.14}$$

Also the (normalized) effective sample size derives from [119]:

$$ESS_{MCMC} = \frac{1}{\tau_{int}} = \frac{1}{1+2\sum_{k=1}^{\infty}\rho(k)}, \tag{2.15}$$

indicating the loss in efficiency due to the usage of a Markov chain. In general, the more correlated the samples are, the less information they contain.

For further introductory information about Monte Carlo methods the reader can consult [117]. For particular information about properties of the transition kernel or about specific MCMC algorithms, such as MALA, Simulated Annealing, Gibbs sampling, we direct the reader to [54, 128, 119]. Whilst MCMC sampling methods are general algorithms that are guaranteed to yield exact estimates in the limit of a large number of samples, the number of required samples for accurate estimates can be too large to make even highly optimized procedures feasible.

2.2.4 Variational Bayes theory

In the last subsections, we discussed different methods to approximate desired probability distributions, such as single point estimates (ML and MAP), local approximation (Laplace approximation) and Monte Carlo techniques. Point estimates and local approximations can be inaccurate whereas Monte Carlo techniques can be computationally expensive. An alternative, on which we focus within this thesis, is *Variational Bayesian* (VB) methods (also known as ensemble learning or variational free energy minimization). VB approximates, for example, the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$ using a simpler distribution $q(\boldsymbol{\Psi})$. VB is more general than a Laplace approximation but computationally much more effective than Monte Carlo methods. This places VB between Laplace approximations and Monte Carlo methods in both, accuracy and computational cost, as is illustrated in Figure 2.2.

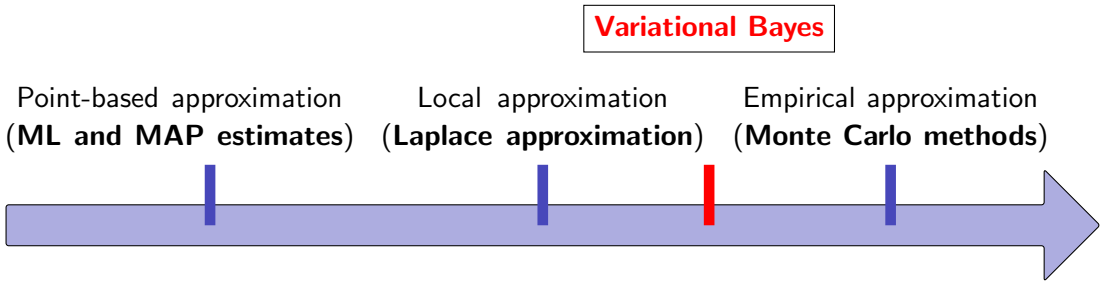


Figure 2.2: Increase of approximation quality and computational cost.

Variational Bayes performs approximate inference by solving an *optimization problem over a family of appropriately selected probability densities* with the objective of minimizing the Kullback-Leibler divergence with the exact posterior. This thesis is devoted to Variational Bayes. A new and fully Bayesian reduction of dimensionality and an extension to multimodal densities will be incorporated, among other things, to accurately solve high-dimensional problems. Before we take a detailed look into

the proposed framework in Chapters 3- 5, we shortly review Variational Bayes in the following.

Variational Bayes has its roots in *calculus of variations*. In calculus of variations, one searches for a function which optimizes a functional. It has its origin in the 18^{th} century and is based on the work of Euler, Lagrange and others. We refer to a historic overview on, e.g., 'Elementa calculi variationum' in [133]. Within the *Variational Bayesian* method functionals are optimized, as in calculus of variations, which explains the relation to its name. The variational lower bound and the Kullback-Leibler divergence are functionals, i.e., they take functions as input arguments and are optimized by searching for an optimal function (here probability density function). Variational Bayesian methods for probabilistic models were introduced in the 1990s by the authors [134, 135] and have their origin in previous work in statistical physics.

Within Variational inference, one approximates the desired probability distribution, here the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$, by a simpler distribution $q(\boldsymbol{\Psi})$:

$$q(\boldsymbol{\Psi}) \approx p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}). \tag{2.16}$$

By employing Jensen's inequality [136, 137], one can construct a variational lower bound $\mathcal{F}(q(\boldsymbol{\Psi}))$ to the log-evidence

$$\begin{aligned} \log p(\hat{\boldsymbol{y}}) &= \log \int p(\boldsymbol{\Psi}, \hat{\boldsymbol{y}})\, d\boldsymbol{\Psi} \\ &= \log \int q(\boldsymbol{\Psi}) \tfrac{p(\boldsymbol{\Psi},\hat{\boldsymbol{y}})}{q(\boldsymbol{\Psi})}\, d\boldsymbol{\Psi} \\ &\geq \int q(\boldsymbol{\Psi}) \log \tfrac{p(\boldsymbol{\Psi},\hat{\boldsymbol{y}})}{q(\boldsymbol{\Psi})}\, d\boldsymbol{\Psi} \\ &=< \log \tfrac{p(\boldsymbol{\Psi},\hat{\boldsymbol{y}})}{q(\boldsymbol{\Psi})} >_{q(\boldsymbol{\Psi})} \\ &= \mathcal{F}(q(\boldsymbol{\Psi})). \end{aligned} \tag{2.17}$$

The lower bound $\mathcal{F}$ has very close connection to the Kullback-Leibler divergence (between approximated posterior and the exact posterior) and the log-evidence [54]:

$$KL(q(\boldsymbol{\Psi})||p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})) = \log p(\hat{\boldsymbol{y}}) - \mathcal{F}(q(\boldsymbol{\Psi})). \tag{2.18}$$

Ideally, one would minimize the KL divergence with respect to $q(\boldsymbol{\Psi})$, but this is not possible as the true posterior is not known. However, as $\log p(\hat{\boldsymbol{y}})$ is constant with respect to $q(\boldsymbol{\Psi})$ and as the KL-divergence is strictly non-negative, maximizing $\mathcal{F}(q(\boldsymbol{\Psi}))$ is equivalent to minimizing $KL(q(\boldsymbol{\Psi})||p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}}))$ with respect to $q(\boldsymbol{\Psi})$, see also Figure 2.3. Convergence to a local maximum of $\mathcal{F}$ is guaranteed due to the fact that the KL divergence is convex and $\mathcal{F}$ is consequently concave [54]. If we allow any distribution for $q(\boldsymbol{\Psi})$, the lower bound is maximal for $q(\boldsymbol{\Psi}) = p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$. However, for complex posteriors this is intractable. Therefore, one usually considers a restricted family of distributions for which the KL divergence is then minimized. Based on the application the restriction is a balance between tractability and accuracy.

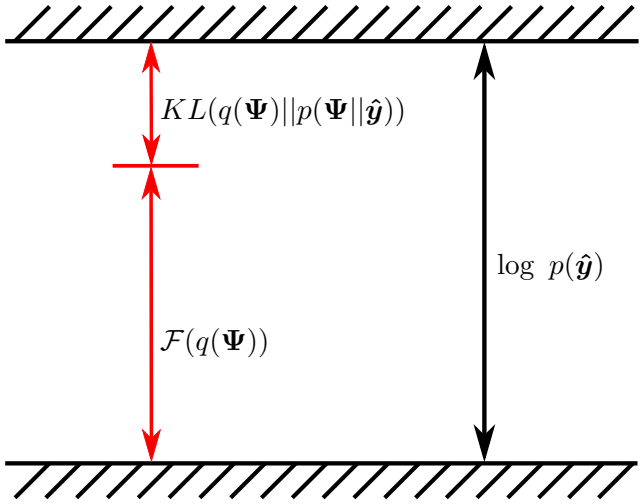


Figure 2.3: Maximizing $\mathcal{F}(q(\mathbf{\Psi}))$ is equivalent to minimizing $KL(q(\mathbf{\Psi})||p(\mathbf{\Psi}|\hat{\boldsymbol{y}}))$ with respect to $q(\mathbf{\Psi})$ because the sum of both terms, $\log p(\hat{\boldsymbol{y}})$, does not depend on $q(\mathbf{\Psi})$ and the KL divergence cannot get negative.

A common and simplifying restriction is to partition $q(\mathbf{\Psi})$ in disjoint groups. This factorized form of the densities corresponds to the *mean field* approximation. The mean field approximation has its origin in statistical physics [138, 139]. It assumes that the posterior distribution of the parameters separated in subgroups are independent:

$$q(\mathbf{\Psi}) \approx \prod_{i=1}^{N} q_i(\mathbf{\Psi}_i). \tag{2.19}$$

The $q_i()$'s can, but do not have to, be of the same kind of probability distribution, e.g., Gaussian or Gamma distributed. When N is the size of $\mathbf{\Psi}$, it is fully factorized, otherwise it is referred to as structured. A complete factorization is usually not required and a *structured mean field approximation* is preferred. This is usually based on the logical and physical appearance in the model. Including the (structured) mean field approximation, the lower bound follows

$$\mathcal{F} \propto \ < \log p(\mathbf{\Psi}, \hat{\boldsymbol{y}}) >_{q(\mathbf{\Psi})} - \sum_{i=1}^{N} < \log q(\mathbf{\Psi}_i) >_{q(\mathbf{\Psi}_i)} . \tag{2.20}$$

The optimal values $q^{opt}(\mathbf{\Psi}_i)$ can be found by differentiating $\mathcal{F}$ with respect to each $q(\mathbf{\Psi}_i)$ and results in:

$$\log q^{opt}(\mathbf{\Psi}_i) = \ < \log p(\mathbf{\Psi}, \hat{\boldsymbol{y}}) >_{\prod_{j \neq i} q(\mathbf{\Psi}_j)}, \tag{2.21}$$

where $< . >_{\prod_{j \neq i} q(\mathbf{\Psi}_j)}$ refers to an expectation with respect to $q(\mathbf{\Psi}_j)$ for all j, except of $j = i$. The following characteristics apply for VB:

- Since $\log q^{opt}(\mathbf{\Psi}_i)$ depends on all values $q^{opt}(\mathbf{\Psi}_j)$ for $j \neq i$, the distributions need to be updated self-consistently.
- Unlike the Laplace approximation, it is not restricted to approximate the posterior by a Gaussian.
- Convergence to a local maximum of $\mathcal{F}$ is guaranteed (as KL is convex). $\mathcal{F}$ is monotonically increasing (until convergence) as the individual updates are concave with respect to each $q(\mathbf{\Psi}_i)$ [54].

2.3 Fundamentals of Continuum and Computational Mechanics

Our application of interest is elastography which is governed by the fundamentals of solid mechanics. For that reason, this subsection contains the basics of nonlinear continuum mechanics and of finite element methods. The forward problem, taking a model and model parameters as inputs and calculating what the observed values should be, is formulated in the general case for a solid mechanics problem and valid for nonlinear material behavior and large deformations. Nevertheless, the solid mechanics problems considered in this work are quasi-static and therefore any time-dependent terms are neglected. For more details on computational solid mechanics we refer the reader to [140, 141, 142].

2.3.1 Deformation, strain and stress

Let the physical domain be described by Ω_0 in $\mathbb{R}^3$ in the reference configuration. The coordinates of the material particles in the undeformed configuration is denoted by $\mathbf{X}$ (material or Lagrangean description) and by $\boldsymbol{x}$ in the deformed configuration (spatial or Eulerian description). The deformation map $\boldsymbol{\phi}$ maps the coordinates of a material point in the reference and physical domain to the spatial configuration in the deformed and spatial domain $\Omega \in \mathbb{R}^3$ (in the static case):

$$\boldsymbol{\phi} : \begin{cases} \Omega_0 \to \Omega \\ \mathbf{X} \to \boldsymbol{x} = \boldsymbol{\phi}(\boldsymbol{X}). \end{cases} \tag{2.22}$$

From here the displacement field $\boldsymbol{u}$, the difference between the spatial and material configurations, follows:

$$\boldsymbol{u}(\boldsymbol{X}) = \boldsymbol{x} - \boldsymbol{X} = \phi(\boldsymbol{X}) - \boldsymbol{X}, \tag{2.23}$$

and the *deformation gradient* $\boldsymbol{F}$ is defined as

$$\boldsymbol{F} = \frac{\partial \phi}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{X}} = \mathbf{1} + \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{X}}. \tag{2.24}$$

The determinant of $\boldsymbol{F}$, $J = det(\boldsymbol{F}) > 0$, marks the change of the volume ($det(\boldsymbol{F}) > 0$ as the volume has to remain positive under deformation) and is for an incompressible material equal to one. To describe the deformation, we use the *Green-Lagrange strain tensor* defined as:

$$\boldsymbol{E} = \frac{1}{2}(\boldsymbol{F}^T\boldsymbol{F} - \boldsymbol{I}). \tag{2.25}$$

The symmetric, positive *right Cauchy-Green tensor* is defined as $\boldsymbol{C} = \boldsymbol{F}^T\boldsymbol{F}$ and is invariant under any superimposed rigid body motion. Related to the material configuration, it is commonly used as a deformation measure. The deformation between a material element and its neighboring elements results in stresses (which is a measure of the force per unit area). By bisecting the body, splitting the body by an imaginary cut, the internal traction force $\boldsymbol{t}$ can be obtained. The traction depends on the orientation of the cut. Nevertheless, the *Cauchy stress tensor* $\tilde{\boldsymbol{\sigma}}$ is independent of this orientation: $\boldsymbol{t} = \tilde{\boldsymbol{\sigma}} \cdot \boldsymbol{n}$, with $\boldsymbol{n}$ being the outward unit normal vector of the plane. The Cauchy stress tensor is the actual stress and the stress defined in material configuration is the *second Piola-Kirchoff stress tensor*, designated as $\boldsymbol{S} = J\boldsymbol{F}^{-1}\tilde{\boldsymbol{\sigma}}\boldsymbol{F}^{-T}$. The internal traction force in the reference configuration is $\boldsymbol{T} = (\boldsymbol{F}\boldsymbol{S}) \cdot \boldsymbol{N}$, where $\boldsymbol{N}$ is the outward normal and $\boldsymbol{T}$ the traction forces in the reference configuration.

2.3.2 Conservation of linear momentum and constitutive law

In the previous subsection, we reviewed one of the governing equations, the *strain-displacement relation*, see Equation (2.25). Another important equation is the *conservation of linear momentum*:

$$\bigtriangledown \cdot (\boldsymbol{F}\boldsymbol{S}) + \rho_0 \boldsymbol{b} = \mathbf{0} \quad \text{in} \quad \Omega_0, \tag{2.26}$$

where $\boldsymbol{b}$ is body force vector (per unit mass) and ρ_0 is the initial density. The governing equations are supplemented by appropriate Dirichlet and Neumann boundary conditions as

$$\boldsymbol{u} = \boldsymbol{u}_b \quad \text{on} \quad \Gamma_u \quad \text{and} \tag{2.27}$$

$$\boldsymbol{F}\boldsymbol{S} \cdot \boldsymbol{N} = \hat{\boldsymbol{T}}^1 \quad \text{on} \quad \Gamma_S. \tag{2.28}$$

Γ_u and Γ_S are subsets of the boundary $\Gamma_0 = \partial\Omega_0$, on which displacement and traction boundary data, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{T}}$, respectively, are specified.
The third governing equation is the *constitutive law*, the relationship between stresses and strains. Many biological materials can be modeled by hyperelastic materials (purely elastic behavior which depends on the current deformation). For hyperelastic materials a *strain energy density function* $w(\boldsymbol{E};\boldsymbol{\phi})$ exists and depends on the invariants of the Lagrangian strain tensor $\boldsymbol{E}$ and the constitutive material parameters $\boldsymbol{\phi}(\boldsymbol{X})$. The stress-strain equation is as follows:

$$\boldsymbol{S} = \frac{\partial w}{\partial \boldsymbol{E}} = \boldsymbol{S}(\boldsymbol{E};\boldsymbol{\phi}). \tag{2.29}$$

The aforementioned governing equations should be complemented with any additional information about the problem or the material, such as incompressibility. In fact, incompressibility is frequently encountered in bio-materials and corresponds to the condition of $det(\boldsymbol{F}) = 1$ at all points in the problem domain.

2.3.3 Numerical formulation and solution of the forward problem

The presented governing equations cannot analytically be solved for the vast majority of problems and one resorts to numerical techniques which discretize the equations and the associated fields. The most prominent approach is the finite element method (FEM), which is employed in this study as well. In the first step, the *weak* form of the partial differential equations is derived. By choosing arbitrary weighting functions $\boldsymbol{v}$ and integrating the residuals of the Equations (2.26),(2.28) and (2.27) over the respective domain we get:

$$\int_{\Omega_0} (\bigtriangledown \cdot (\boldsymbol{F}\boldsymbol{S}) + \rho_0 \boldsymbol{b}) \cdot \boldsymbol{v}\ dV_0 + \int_{\Gamma_S} (\hat{\boldsymbol{T}} - \boldsymbol{F}\boldsymbol{S} \cdot \boldsymbol{N}) \cdot \boldsymbol{v}\ dA_0 = 0, \tag{2.30}$$

such that $\boldsymbol{v} = \boldsymbol{0}$ on Γ_u. Applying the Gauss divergence theorem it follows:

$$\int_{\Omega_0} \boldsymbol{F}\boldsymbol{S} : (\bigtriangledown \boldsymbol{v})^T\ dV_0 = \int_{\Gamma_S} \hat{\boldsymbol{T}} \cdot \boldsymbol{v}\ dA_0 + \int_{\Omega_0} \rho_0 \boldsymbol{b} \cdot \boldsymbol{v}\ dV_0, \tag{2.31}$$

which can equally be derived by the principle of virtual work, where the weighting functions can be seen as virtual displacements. Subsequently, the problem domain can

[1]When $\hat{\boldsymbol{T}}$ is not known in the material configuration but in the deformed configuration $\hat{\boldsymbol{t}}$, the corresponding formulations in the deformed configuration should be used. Prestressing techniques can alternatively be applied for working in the reference configuration with a loaded stress configuration [143].

be discretized into finite elements (FE) in space by subdividing the domain into d_{FE} non-overlapping subdomains

$$\Omega_0 \approx \bigcup_{e=1}^{d_{FE}} \Omega_0^e. \tag{2.32}$$

Shape functions are used for the interpolation of the unknown fields. Since this is a very mature subject from a theoretical and computational point of view, we do not provide further detail here but point the interested reader to one of many available books [144, 145]. More specifically in the context of inverse problems for (in)compressible elasticity we refer to [43, 146].

Most often, all unknowns of the forward problem are expressed in terms of the discretized displacement field, which is here designated by $\boldsymbol{U} \in \mathbb{R}^n$. An approximate solution of the forward problem can be found by solving an n-dimensional system of nonlinear algebraic equations which can be written in residual form as:

$$\boldsymbol{r}(\boldsymbol{U}; \boldsymbol{\Psi}) = \boldsymbol{0}. \tag{2.33}$$

We denote the residual by $\boldsymbol{r} : \mathbb{R}^n \times \mathbb{R}^{d_\Psi} \to \mathbb{R}^n$ and the *discretized* vector of the constitutive material parameters $\phi(\boldsymbol{X})$ by $\boldsymbol{\Psi} \in \mathbb{R}^{d_\Psi}$. The system can be discretized in many different ways. For example, the same shape and weighting functions can be adopted (Bubnov-Galerkin method). Then each entry of the vector $\boldsymbol{\Psi}$ corresponds to the value of the material parameter at a specific nodal point. Frequently it is assumed that the value of the constitutive parameters is constant within each finite element. In this case d_Ψ coincides with the number of elements d_{FE} in the FE mesh. We would like to point out that the discretization of $\boldsymbol{\Psi}$ does not need to be associated with the discretization used for the governing equations and a finer or coarser discretization might be employed. However, if the material properties exhibit significant variability within each finite element, i.e., if $d_\Psi \gg n$, special care has to be taken in formulating the finite element solution and multiscale schemes might need to be employed [147].

The resulting forward problem refers to Equation (2.33), with given discretized material parameters $\boldsymbol{\Psi}$. For most nonlinear cases the calculation of $\boldsymbol{U}$ by directly solving Equation (2.33) cannot be done and an iterative approach, e.g., the Newton-Raphson method, needs to be applied to solve the forward problem. The Newton-Raphson method [148, 149, 142], also called Newton's method, finds the root of a real-valued differentiable function, here the residual, with respect to $\boldsymbol{U}$. For this purpose $\boldsymbol{r}(\boldsymbol{U}; \boldsymbol{\Psi})$ is linearized around $\boldsymbol{U}^{(k)}$, where k is the iteration number. The root of the function is then approximated by the root of the linearized function. Then, the derived approximation of $\boldsymbol{U}^{(k)}$ is used for the next linearization. Summarized, the displacements are found iteratively by:

$$\boldsymbol{U}^{(k+1)} = \boldsymbol{U}^{(k)} - \boldsymbol{K}^{-1}(\boldsymbol{U}^{(k)}) \boldsymbol{r}(\boldsymbol{U}^{(k)}; \boldsymbol{\Psi}), \tag{2.34}$$

where $\boldsymbol{K}$ is the tangent stiffness matrix

$$\boldsymbol{K}(\boldsymbol{U}^{(k)}) = \frac{\partial \boldsymbol{r}(\boldsymbol{U}^{(k)}; \boldsymbol{\Psi})}{\partial \boldsymbol{U}^{(k)}}. \tag{2.35}$$

The iterative procedure is repeated until a convergence criterion is met, e.g., the Euclidean norm of the residual is smaller than a certain threshold $||\boldsymbol{r}(\boldsymbol{U}^{(k)}; \boldsymbol{\Psi})|| < tol$. The speed of convergence depends on the initial estimate $\boldsymbol{U}^{(k=0)}$. The Newton-Raphson method is a locally convergent scheme and when a stationary point is encountered the algorithm will be terminated based on a zero tangent stiffness matrix.

The Newton-Raphson method requires the solution of a linearized system of equations in each iteration. This includes each time the formation and inversion of the tangent stiffness matrix as well as the evaluation of the residual. Alternatively, approximations of the true tangent stiffness matrix and/or its inverse have been established. A quasi-Newton method, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method is used within this thesis. We refer to [150, 151] for more details.

We will later compare displacements obtained by solving the forward model with measured ones. Often the experimental measurements/ observations are (noisy) displacements at specific locations in the physical domain. As a result, one is interested in a subset or in a lower-dimensional function of $\boldsymbol{U}$ at the same locations as the observations. We denote these displacements by $\boldsymbol{y} \in \mathbb{R}^{d_y}$ and they can be formally expressed as $\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{U}$, where $\boldsymbol{Q}$ is a Boolean matrix which selects the entries of interest from $\boldsymbol{U}$. Since $\boldsymbol{U}$ depends on $\boldsymbol{\Psi}$, $\boldsymbol{y}$ is also a function of $\boldsymbol{\Psi}$, i.e., $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{\Psi})$. We emphasize that this function is generally *highly nonlinear*.

2.3.4 Inverse problem and adjoint formulation

In the previous subsection, we revised the ingredients of the solution of the forward problem. In our application of interest, elastography, it relates to the derivation of the displacements of the material given the material parameters and boundary conditions. Despite this, we are interested in identifying the material parameters $\boldsymbol{\Psi}$ based on observed displacements $\hat{\boldsymbol{y}} \in \mathbb{R}_{d_y}$. This refers to an *inverse problem* as depicted in Figure 2.4. Inverse problems are of great interest in many disciplines as they inform us about underlying model parameters which cannot be observed directly. Detailed information about solution strategies are found in [45, 131, 152].

To solve the inference scheme of the inverse problem proposed later, we need both the solution vector of the forward problem $\boldsymbol{U}(\boldsymbol{\Psi})$ and the derivatives $\frac{\partial \boldsymbol{y}(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}}$, as in Equations 3.15,4.20. The computation of the derivatives of the response with respect to model parameters is a well-studied subject in the context of PDE-constrained optimization.

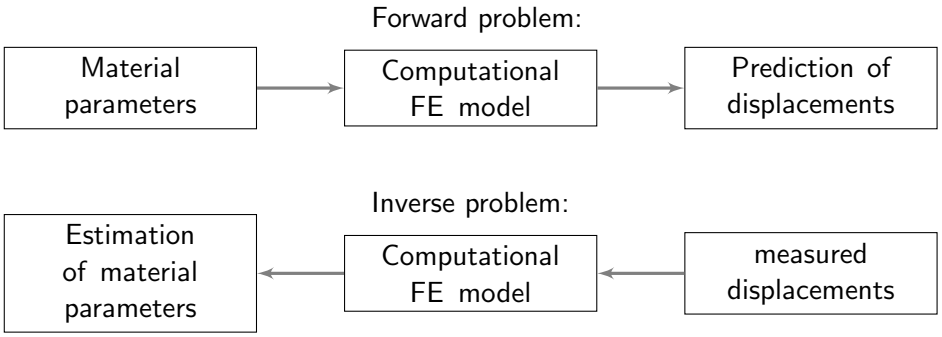


Figure 2.4: Presentation of the forward and inverse problem in terms of the application elastography.

The inverse problem can be formulated as: find the spatial distribution of the material parameters $\boldsymbol{\Psi}$ that minimizes the objective function f:

$$f(\boldsymbol{U}) = \frac{1}{2}||\boldsymbol{Q}\boldsymbol{U}(\boldsymbol{\Psi}) - \hat{\boldsymbol{y}}||^2. \tag{2.36}$$

For the computation of the gradient of the objective function the adjoint approach is used widely [153, 154, 155].

For any scalar function $f(\boldsymbol{U})$, one can employ the adjoint form of Equation (2.33), according to which:

$$\frac{df}{d\Psi_k} = -\nu_i \frac{\partial r_i}{\partial \Psi_k}, \tag{2.37}$$

where the adjoint variable $\boldsymbol{\nu} \in \mathbb{R}^n$ is defined such as:

$$\nu_j \frac{\partial r_j}{\partial U_i} = \frac{\partial f}{\partial U_i} \quad \text{or} \quad \boldsymbol{J}^T \boldsymbol{\nu} = \frac{\partial f}{\partial \boldsymbol{U}}. \tag{2.38}$$

We note that $\frac{\partial r_j}{\partial U_i}$ is the Jacobian of the residuals in Equation (2.33), evaluated at the solution $\boldsymbol{U}(\boldsymbol{\Psi})$. We point out that if a direct solver for the solution of the linear system in Equation (2.33) is employed, then the additional cost of evaluating $\frac{df}{d\boldsymbol{\Psi}}$ is minimal as the Jacobian would not need to be re-factorized for solving Equation (2.38).[2] In the context of the problems considered in this thesis repeated use of Equation (2.38) is made, where f is a different component of the observables. As such the overall cost increases proportionally with the number of observables (displacements in our problems) that are available. In problems where n is so large that it precludes the use of direct solvers the cost of the solution of the adjoint equations can be increased. Nevertheless, it is comparable to the cost of a forward solution.

[2]The cost of evaluating $\frac{\partial r_i}{\partial \Psi_k}$ is negligible compared to other terms as it scales linearly with the number of elements/nodes.

In cases where both n and the dimension of $\mathbf{\Psi}$ are high, advanced iterative solvers, suitable for multiple right-hand sides, must be employed [156, 157]. These imply an added computational burden which scales sublinearly with the dimension of $\mathbf{\Psi}$.

In the incompressible case of the solid mechanics problem, pressure must be taken into account. For that purpose the pressure trial solutions $p \in L_2(\Omega_0)$ and weighting functions $q \in L_2(\Omega_0)$ should also be introduced [146].

2.4 Outlook

In the previous subsection, the fundamentals of continuum and numerical mechanics are outlined and used to build a forward problem. Afterwards, an inverse problem is formulated to resolve unknown quantities based on observations.

Although many different elastography techniques exist, underlying uncertainties, for example, introduced by noise and incomplete observations, are usually neglected. Recently published research incorporates the underlying uncertainties. A Bayesian estimation of material parameters is included to model a non-rigid image registration more accurately [111]. However, only six material parameters and their uncertainties are derived, assuming constant properties over different regions. In [114, 108] more material parameters and their uncertainties are estimated for a linear elastic material model. Nevertheless, the unknowns are derived by sampling, which is computationally very expensive for many unknowns.

In the following chapter, we develop an uncertainty quantification method to solve high-dimensional nonlinear inverse problems. Then, within the application of strain elastography - usually a high-dimensional problem - we not only quantify the underlying mechanical properties but furthermore account for their uncertainties. We especially focus on reducing the number of dimensions as well as computational cost in order to make derived solution strategies achievable for large systems. In addition, we propose methods for quantifying model errors. Although we use elastography as our application example, these methods are easily transferable to other problems of interest.

Chapter 3

Sparse Variational Bayesian approximations for nonlinear inverse problems

> *[...] the statistician knows [...] that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world.*

George Box, 1919-2013 [158].

This chapter is based on the publication: I. M. Franck, P.S. Koutsourelakis, *Sparse Variational Bayesian approximations for nonlinear inverse problems: Applications in nonlinear elastography*, Computer Methods in Applied Mechanics and Engineering, Volume 299, 1 February 2016, Pages 215-244 [159]

3.1 Problem description and introduction

Inference methods, such as Monte Carlo or Variational Bayesian strategies, become usually problematic with an increasing number of unknowns and problem-dimensionality. In such problems, dimensionality reduction plays a pivotal role, more specifically on the identification of lower-dimensional features that provide the strongest signature to the unknowns and the corresponding posterior. Discovering a sparse set of features has attracted great interest in many applications, such as in the representation of natural images [160] or more generally in signal processing applications. A host of algorithms have been developed for finding such representations and also appropriate dictionaries for achieving this goal [161, 162, 163, 164]. While all these tools are pertinent to the present problem they differ in a fundamental way. They are based on several data/observations/instantiations of the vector that we seek to represent. However, in our problems we do not have such direct observations, i.e., the available data pertains to the output of a model which is nonlinearly and implicitly dependent on the vector of unknowns. Furthermore, we are primarily interested in approximating the posterior of this vector rather than simply performing dimensionality reduction. *We demonstrate how this can be done by using a fully Bayesian formulation and employing the marginal likelihood or evidence as the ultimate model validation metric for any proposed dimensionality reduction.*

Let the vector $\boldsymbol{\Psi} \in \mathbb{R}^{d_\Psi}$ represent any model parameters for which a model output $\boldsymbol{y}(\boldsymbol{\Psi}) \in \mathbb{R}^{d_y}$ is available (forward run) and the calibration of the model is of interest. We also presuppose the availability of the derivatives with respect to the model parameters $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}$. For problems of practical interest, it is assumed that the dimension d_Ψ of the unknowns is very large which poses a significant hindrance in finding proper regularization (in deterministic settings [165]) or in specifying appropriate priors (in probabilistic settings [166, 167]). The primary focus of the Bayesian model developed in this section is two-fold:

- Find lower-dimensional representations of the unknown parameter vector $\boldsymbol{\Psi}$ that capture as much as possible of the associated posterior density.

- Enable the computation of the posterior density with as few forward calls (i.e., evaluations of $\boldsymbol{y}(\boldsymbol{\Psi}), \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}$) as possible.

We denote $\hat{\boldsymbol{y}} \in \mathbb{R}^{d_y}$ the vector of observations/measurements. In the context of elastography the observations are displacements (in the static case) and/or velocities (in dynamics). The extraction of this data from images (ultrasound or MRI) is a challenging topic that requires sophisticated image registration techniques [168, 169]. Naturally, this compromises the informational content of the raw data (i.e., the images). In this study, we ignore the error introduced by the image registration process, as the emphasis

is on the inversion of the continuum mechanics, PDE-based model and assume that the displacement data are contaminated with noise.

We postulate the presence of i.i.d. Gaussian noise, denoted here by the random vector $\boldsymbol{z} \in \mathbb{R}^{d_y}$, such that:

$$\hat{\boldsymbol{y}} = \boldsymbol{y}(\boldsymbol{\Psi}) + \boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{I}_{d_y}). \tag{3.1}$$

$\mathcal{N}(\boldsymbol{z}|\mathbf{0}, \tau^{-1}\mathbf{I}_{d_y})$ denotes a multivariate normal distribution of $\boldsymbol{z}$ with a mean $\mathbf{0}$ and a covariance of $\tau^{-1}\mathbf{I}_{d_y}$. Often a short-hand notation, skipping for a simplified notation the random variable which obeys the normal distribution, is used: $\mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{I}_{d_y})$. We assume that each entry of $\boldsymbol{z}$ has zero mean and an unknown variance τ^{-1}, which will also be inferred from the data. We note that other models can also be employed, such as impulsive noise to account for outliers due to instrument calibration (e.g., to account for faulty sensors) or experimental conditions [170]. Generally, the difference between observed and model-predicted outputs can be attributed not only to observation errors (noise), but also to model discrepancies arising from the discretization of the governing equations. Another source of error can be an inadequacy of the model, which captures the underlying physical process, itself. While the former source can be reduced by considering very fine discretizations (at the cost of increasing the dimensionality of the state vector $\boldsymbol{u}$ and potentially $\boldsymbol{\Psi}$), the latter requires a much more thorough treatment [46, 171, 172, 100, 108, 173, 174], on which we focus on in Chapter 5. Within this chapter such model errors are lumped with observation errors in the $\boldsymbol{z}$-term.

The likelihood function of the observed data $\hat{\boldsymbol{y}}$, i.e., its conditional probability density given the model parameters $\boldsymbol{\Psi}$ (and implicitly the model $\mathcal{M}$ itself, as described by Equation (2.33) and the resulting $\boldsymbol{y}(\boldsymbol{\Psi})$) and τ is:

$$p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi}, \tau) = \left(\frac{\tau}{2\pi}\right)^{d_y/2} e^{-\frac{\tau}{2}||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\Psi})||^2}. \tag{3.2}$$

In the Bayesian advocated framework, one also needs to specify priors on the unknown parameters. We defer a detailed discussion of the priors associated with $\boldsymbol{\Psi}$ for the next section where the dimensionality reduction aspects are discussed. With regard to the noise precision τ we employ a (conditionally) conjugate Gamma prior [175], i.e.,

$$p_\tau(\tau) = Gamma(a_0, b_0). \tag{3.3}$$

The values of the parameters are taken $a_0 = b_0 = 0$ in the following examples. This corresponds to a limiting case where the density degenerates to an improper, non-informative Jeffreys prior, i.e., $p_\tau(\tau) \propto \frac{1}{\tau}$ that is scale invariant [48]. Naturally, more informative choices can be made if such information is available a priori.

3.2 Methods

3.2.1 Dimensionality reduction for $\boldsymbol{\Psi}$

One way to enforce dimensionality reduction is by an appropriate prior specification. For example, in [176], the Fourier transform coefficients of $\boldsymbol{\Psi}$ corresponding to small-wavelength fluctuations were turned-off by assigning zero prior probability to non-zero values. While such an approach achieves the goal of dimensionality reduction it does not take into account the forward model in doing so. The nonlinear map $\boldsymbol{y}(\boldsymbol{\Psi})$ as well as the available data $\hat{\boldsymbol{y}}$ provide varying amounts of information for identifying different features of $\boldsymbol{\Psi}$. One would expect the likelihood (which measures the degree of fit of model predictions with the data) to exhibit different levels of sensitivity along different directions in the $\boldsymbol{\Psi}$-space. Consider the Laplace's method for example, which is based on a semi-analytic Gaussian approximation around the Maximum-A-Posteriori estimate $\boldsymbol{\Psi}_{MAP}$ (Section 2.2.2). The negative of the Hessian of the log-posterior (assuming this is positive-definite) serves as the covariance matrix. As it was shown in [74], in many inverse problems this covariance matrix exhibits a significant discrepancy in its eigenvalues which was exploited in constructing low-rank approximations. At one extreme, there would be principal directions (with small variance) along which a small distance from the location of $\boldsymbol{\Psi}_{MAP}$ would cause a huge decrease in the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$ and on the other, there would principal directions (with large variance) along which the posterior would remain almost constant. Such principal directions will naturally encapsulate the effect of the log-prior. In the proposed scheme however, *only* the data log-likelihood affects the directions with the maximal posterior variance [75]. Perhaps more importantly, we propose a unified framework where the identification of the subspace with the largest posterior variance is performed *simultaneously* with the inference of the posterior under the same Variational Bayesian objective. This yields not only a highly efficient algorithm (in terms of the number of forward solves) but also a highly extendable framework as discussed in the conclusion of this chapter.

The inference and dimensionality reduction problems are approached by employing a fully Bayesian formulation and invoking the quality of the approximation to the posterior as our guiding objective. To that end, we postulate the following representation for the high-dimensional vector of unknowns $\boldsymbol{\Psi}$:

$$\underbrace{\boldsymbol{\Psi}}_{d_\psi \times 1} = \underbrace{\boldsymbol{\mu}}_{d_\psi \times 1} + \underbrace{\boldsymbol{W}}_{d_\psi \times d_\theta} \underbrace{\boldsymbol{\Theta}}_{d_\theta \times 1} + \underbrace{\boldsymbol{\eta}}_{d_\psi \times 1}. \tag{3.4}$$

The motivation behind such a decomposition is quite intuitive as it resembles a Principal Component Analysis (PCA) model [177]. The vector $\boldsymbol{\mu}$ represents the mean value of the representation of $\boldsymbol{\Psi}$ and the columns of the orthogonal matrix $\boldsymbol{W} \in \mathbb{R}^{d_\Psi \times d_\Theta}$ span the aforementioned subspace with reduced coordinates $\boldsymbol{\Theta} \in \mathbb{R}^{d_\Theta}$. d_Θ is the number of

reduced variables and $\boldsymbol{\eta} \in \mathbb{R}^{d_\Psi}$ captures the residual variance that complements to the main effects.

The linear decomposition of a high-dimensional vector, such as $\boldsymbol{\Psi}$, has received a lot of attention in several different fields. Most commonly $\boldsymbol{\Psi}$ represents a high-dimensional signal (e.g., an image, an audio/video recording) and $\boldsymbol{W}$ consists of an over- or under-complete basis set [160, 178] which attempts to encode the signal as *sparsely* as possible. Significant advances in Compressed Sensing [179] or Sparse Bayesian Learning [180] have been achieved in recent years along these lines. They are based on several observations of $\boldsymbol{\Psi}$, whereas in our problem we do not have such direct observations, i.e., the data available pertains to $\boldsymbol{y}$ which is nonlinearly and implicitly dependent on $\boldsymbol{\Psi}$. Furthermore, we are primarily interested in approximating the posterior on $\boldsymbol{\Psi}$ rather than the dimensionality reduction itself.

We focus now on the representation of Equation (3.4) and proceed to discuss the identification of $\boldsymbol{\mu}$, $\boldsymbol{W}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$. In a fully Bayesian setting these parameters would be equipped with priors, say $p_\mu(\boldsymbol{\mu}), p_W(\boldsymbol{W}), p_\Theta(\boldsymbol{\Theta}), p_\eta(\boldsymbol{\eta})$ respectively, and their *joint* posterior would be sought:

$$p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau | \hat{\boldsymbol{y}}) \propto p(\hat{\boldsymbol{y}} | \boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)\ p_\mu(\boldsymbol{\mu})\ p_W(\boldsymbol{W})\ p_\Theta(\boldsymbol{\Theta})\ p_\eta(\boldsymbol{\eta})\ p_\tau(\tau), \quad (3.5)$$

where $p_\tau(\tau)$ represents the Gamma prior for τ discussed in Equation (3.3). Such an inference problem would in general be formidable, particularly with regard to $\boldsymbol{\mu}$ and $\boldsymbol{W}$ whose dimension is dominated by d_Ψ. To address this difficulty we propose computing point estimates for $\boldsymbol{\mu}$ and $\boldsymbol{W}$ while inferring the whole posterior of $\boldsymbol{\Theta}, \boldsymbol{\eta}$. In computing the point estimates for $\boldsymbol{\mu}$ and $\boldsymbol{W}$, the natural objective function would be the marginal posterior $p(\boldsymbol{\mu}, \boldsymbol{W} | \hat{\boldsymbol{y}})$:

$$p(\boldsymbol{\mu}, \boldsymbol{W} | \hat{\boldsymbol{y}}) = \int p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau | \hat{\boldsymbol{y}})\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta}\ d\tau. \quad (3.6)$$

In such a case the point estimates for $\boldsymbol{\mu}, \boldsymbol{W}$ would be the Maximum-a-Posteriori-Estimates (MAP). We note that (up to an additive constant):

$$\begin{aligned}
&\log p(\boldsymbol{\mu}, \boldsymbol{W} | \hat{\boldsymbol{y}}) = \\
&\quad = \log \textstyle\int p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau | \hat{\boldsymbol{y}})\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta}\ d\tau \\
&\quad = \log \textstyle\int p(\hat{\boldsymbol{y}} | \boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)\ p_\Theta(\boldsymbol{\Theta})\ p_\eta(\boldsymbol{\eta})\ p_\tau(\tau)\ p_\mu(\boldsymbol{\mu})\ p_W(\boldsymbol{W})\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta}\ d\tau \\
&\quad = \log \textstyle\int p(\hat{\boldsymbol{y}} | \boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)\ p_\Theta(\boldsymbol{\Theta})\ p_\eta(\boldsymbol{\eta})\ p_\tau(\tau)\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta}\ d\tau \\
&\quad + \log p_\mu(\boldsymbol{\mu}) + \log p_W(\boldsymbol{W}) \\
&\quad = \log \textstyle\int \left(\frac{\tau}{2\pi}\right)^{d_y/2} e^{-\frac{\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2}\ p_\Theta(\boldsymbol{\Theta})\ p_\eta(\boldsymbol{\eta})\ p_\tau(\tau)\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta}\ d\tau \\
&\quad + \log p_\mu(\boldsymbol{\mu}) + \log p_W(\boldsymbol{W}).
\end{aligned} \quad (3.7)$$

We indicate that such an integration is analytically impossible primarily due to the nonlinear and implicit nature of $\boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})$ and secondarily due to the coupling

of $\boldsymbol{\Theta}, \boldsymbol{\eta}$ and τ. To that end, we employ a Variational Bayesian approximation [54] to the integral in Equation (3.7). We provide further details in the next subsection. We mention that similar approximations have been employed in previous works [84, 170, 83] in order to expedite Bayesian inference. The novel element of this work pertains to the dimensionality reduction that can be achieved.

3.2.2 Variational Bayesian Expectation Maximization algorithm

In practice the posterior is often referred to as analytically intractable. Sampling methods, such as MCMC, or approximation methods, such as Laplace approximation, are often either too expensive or not accurate enough (Section 2.2). The option we discuss in the following is the *Variational Bayes* method, introduced in Section 2.2.4, which is now extended to incorporate dimensionality reduction.

Consider an arbitrary joint density $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ on the latent variables $\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau$. Then by employing Jensen's inequality, one can construct a lower bound to the log-marginal-posterior $\log p(\boldsymbol{\mu}, \boldsymbol{W}|\hat{\boldsymbol{y}})$ of Equation (3.7) as follows:

$$\begin{aligned} \log p(\boldsymbol{\mu}, \boldsymbol{W}|\hat{\boldsymbol{y}}) &= \log \int p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}})\, d\boldsymbol{\Theta}\, d\boldsymbol{\eta}\, d\tau \\ &= \log \int q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) \tfrac{p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)}\, d\boldsymbol{\Theta}\, d\boldsymbol{\eta}\, d\tau \\ &\geq \int q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) \log \tfrac{p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)}\, d\boldsymbol{\Theta}\, d\boldsymbol{\eta}\, d\tau \\ &= \mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W}). \end{aligned} \tag{3.8}$$

The Kullback-Leibler divergence between $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ and the (conditional) posterior on $(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$:

$$p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W}) = \frac{p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}})}{p(\boldsymbol{\mu}, \boldsymbol{W}|\hat{\boldsymbol{y}})}, \tag{3.9}$$

relates to the variational lower-bound $\mathcal{F}$:

$$\begin{aligned} KL\left(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)||p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W})\right) &= -\left\langle \log \tfrac{p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \right\rangle_q \\ &= -\left\langle \log \tfrac{p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}})}{p(\boldsymbol{\mu}, \boldsymbol{W}|\hat{\boldsymbol{y}})\ q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \right\rangle_q \\ &= \log p(\boldsymbol{\mu}, \boldsymbol{W}|\hat{\boldsymbol{y}}) - \mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W}), \end{aligned} \tag{3.10}$$

where $< . >_q$ is the expectation with regard to q. If q is not further specified it relates to the full joint density, otherwise to the specified marginal density, e.g., to $q(\boldsymbol{\Theta})$ for $< . >_{\boldsymbol{\Theta}}$. The KL-divergence becomes 0 when $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) \equiv p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W})$ (see Section 2.1). Hence, for a given $\boldsymbol{\mu}, \boldsymbol{W}$, constructing a good approximation to the conditional posterior (in the KL-divergence sense) is equivalent to maximizing the lower bound $\mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W})$ with regard to $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ (Section 2.2.4).

The aforementioned discussion suggests an iterative optimization scheme that resembles the Variational Bayes - Expectation-Maximization (VB-EM) methods that have

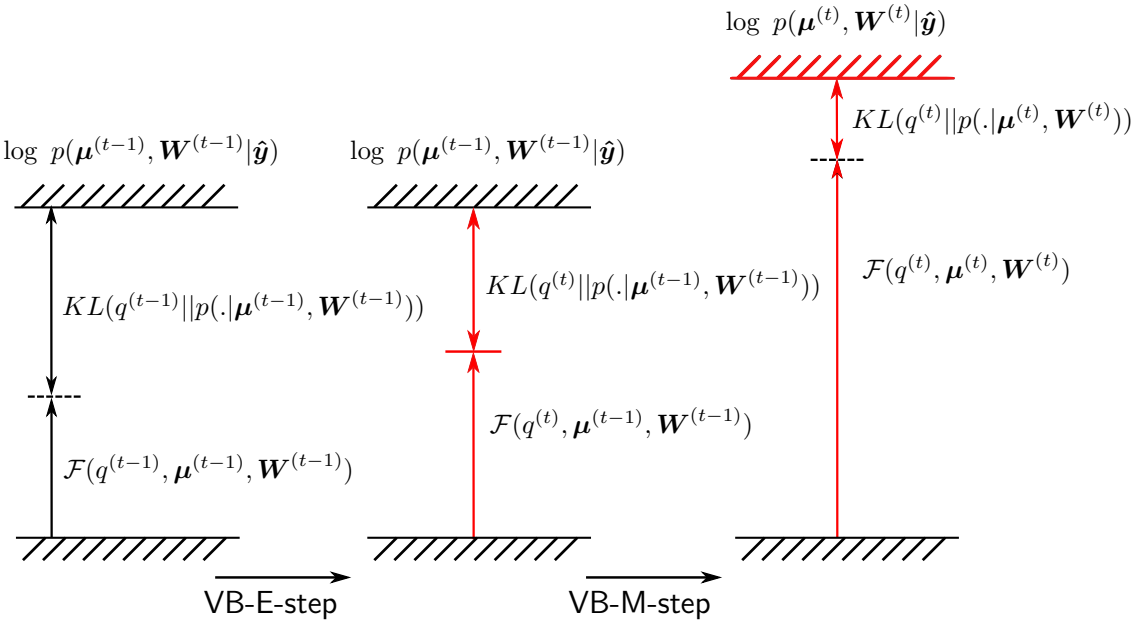


Figure 3.1: During the VB-E step, optimization with respect to the approximating distribution q takes place, whereas during the VB-M step, $\mathcal{F}$ is optimized with respect to the model parameters $\boldsymbol{\mu}, \boldsymbol{W}$ (adapted from [79]).

appeared in Machine Learning literature [79]. At each iteration t, one alternates between (Figure 3.1):

- **VB-Expectation**: Given $(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{W}^{(t-1)})$, find:

$$q^{(t)}(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) = \arg\max_{q} \mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)), \boldsymbol{\mu}^{(t-1)}, \boldsymbol{W}^{(t-1)}), \tag{3.11}$$

- **VB-Maximization**: Given $q^{(t)}(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$, find:

$$(\boldsymbol{\mu}^{(t)}, \boldsymbol{W}^{(t)}) = \arg\max_{\boldsymbol{\mu}, \boldsymbol{W}} \mathcal{F}(q^{(t)}(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W}). \tag{3.12}$$

In plain terms, the strategy advocated in order to carry out the inference task can be described as a generalized coordinate ascent with regard to $\mathcal{F}$ (Figure 3.2).

From Equations (3.5) and (3.8), we have that:

$$\begin{aligned}
&\mathcal{F}\left(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W}\right) = \\
&= \int q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) \log \tfrac{p(\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau|\hat{\boldsymbol{y}})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \; d\boldsymbol{\Theta} \; d\boldsymbol{\eta} \; d\tau \\
&= \int q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) \log \tfrac{p(\hat{\boldsymbol{y}}|\boldsymbol{\mu}, \boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) \; p_{\Theta}(\boldsymbol{\Theta}) \; p_{\eta}(\boldsymbol{\eta}) \; p_{\tau}(\tau)}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \; d\boldsymbol{\Theta} \; d\boldsymbol{\eta} \; d\tau \\
&+ \log p_{\mu}(\boldsymbol{\mu}) + \log p_{W}(\boldsymbol{W}) \\
&= \left\langle \log \tfrac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{\mu}, \boldsymbol{W}) \; p_{\Theta}(\boldsymbol{\Theta}) \; p_{\eta}(\boldsymbol{\eta}) \; p_{\tau}(\tau)}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \right\rangle_{q} + \log p_{\mu}(\boldsymbol{\mu}) + \log p(\boldsymbol{W}) \\
&= \hat{\mathcal{F}}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W}) + \log p_{\mu}(\boldsymbol{\mu}) + \log p_{W}(\boldsymbol{W}),
\end{aligned} \tag{3.13}$$

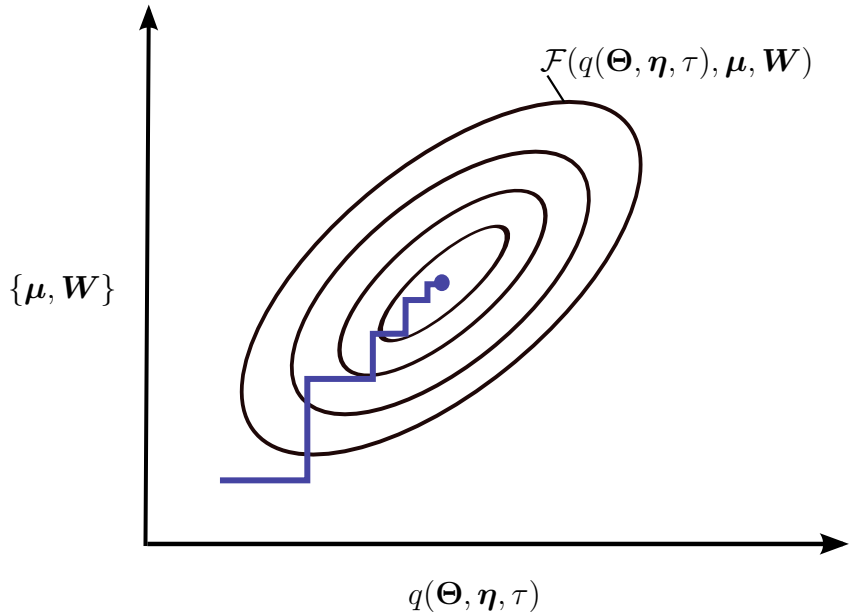


Figure 3.2: Schematic illustration of the advocated Variational Bayesian Expectation-Maximization (VB-EM, [79]).

where (up to an additive constant):

$$\begin{aligned}\hat{\mathcal{F}}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau), \boldsymbol{\mu}, \boldsymbol{W}) &= \left\langle \log \frac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{\mu}, \boldsymbol{W})\ p_\Theta(\boldsymbol{\Theta})\ p_\eta(\boldsymbol{\eta})\ p_\tau(\tau)}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \right\rangle_q \\ &= \left\langle \log \left(\frac{\tau}{2\pi}\right)^{d_y/2} e^{-\frac{\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2} \right\rangle_q \\ &+ \left\langle \log \frac{p_\Theta(\boldsymbol{\Theta})\ p_\eta(\boldsymbol{\eta})\ p_\tau(\tau)}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)} \right\rangle_q .\end{aligned} \tag{3.14}$$

To alleviate the difficulties with the log-likelihood integral above we employ the following approximations:

- We linearize the map $\boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})$ at $\boldsymbol{\mu}$. Hence:

$$\boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta}) = \boldsymbol{y}(\boldsymbol{\mu}) + \boldsymbol{G}(\boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta}) + \mathcal{O}(||\boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta}||^2), \tag{3.15}$$

  where $\boldsymbol{G} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}|_{\boldsymbol{\Psi}=\boldsymbol{\mu}}$ is the gradient of the map at $\boldsymbol{\mu}$.

  By keeping the first order terms from Equation (3.15), the term $||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2$ in the exponent of the likelihood becomes:

$$\begin{aligned}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2 &= ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}) - \boldsymbol{G}\boldsymbol{W}\boldsymbol{\Theta} - \boldsymbol{G}\boldsymbol{\eta}||^2 \\ &= ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 - 2(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}))^T \boldsymbol{G}\boldsymbol{W}\boldsymbol{\Theta} \\ &+ \boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W} : \boldsymbol{\Theta}\boldsymbol{\Theta}^T \\ &- 2\boldsymbol{\eta}^T\boldsymbol{G}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}) - \boldsymbol{G}\boldsymbol{W}\boldsymbol{\Theta}) \\ &+ \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{\eta}.\end{aligned} \tag{3.16}$$

We note here that a quadratic expression with respect to $\boldsymbol{\Theta}$ could also be obtained by considering the 2^{nd} order Taylor series of $||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2$ around $\boldsymbol{\mu}$ directly. In particular, if we denote by $\boldsymbol{g} = \frac{\partial||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\Psi})||^2}{\partial \boldsymbol{\Psi}}|_{\boldsymbol{\Psi}=\boldsymbol{\mu}}$ and $\boldsymbol{H} = \frac{\partial||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\Psi})||^2}{\partial \boldsymbol{\Psi}\partial \boldsymbol{\Psi}^T}|_{\boldsymbol{\Psi}=\boldsymbol{\mu}}$ and keeping only up to second order terms yields:

$$\begin{aligned} ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2 \quad &= ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 + \boldsymbol{g}^T(\boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta}) \\ &+\tfrac{1}{2}\boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W} : \boldsymbol{\Theta}\boldsymbol{\Theta}^T \\ &+\boldsymbol{\eta}^T\boldsymbol{H}\boldsymbol{W}\boldsymbol{\Theta} \\ &+\tfrac{1}{2}\boldsymbol{\eta}^T\boldsymbol{H}\boldsymbol{\eta}. \end{aligned} \tag{3.17}$$

The computation of 2^{nd} order derivatives $\boldsymbol{H}$ can also be addressed within the adjoint framework. We refer the interested reader to [154, 181] as we do not pursue this possibility further in this work. The ensuing expressions are based on Equation (3.16) but can be readily adjusted to include the terms in Equation (3.17) instead[1].

We note that by making use of the linearization of the map $\boldsymbol{y}(\boldsymbol{\Psi})$ and the Variational Bayesian approximation, one can obtain a tractable approximation of the posterior of the latent parameters $\boldsymbol{\Theta}, \boldsymbol{\eta}$ and τ. This will enable us to ultimately identify all model parameters and through this process the optimal subspace for approximating the posterior on $\boldsymbol{\Psi}$. This will be explained in detail when the final algorithm is presented in Section 3.2.4.

- The aforementioned equations for the VB-Expectation step imply that probabilistic inference can be expressed in terms of a parametric optimization problem. One can adopt a functional form for $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ depending on an appropriate set of parameters and identify their optimal value by minimizing the KL-divergence with the posterior or equivalently maximizing $\mathcal{F}$. We adopt a *structured mean-field* approximation (see Equation (2.19)) where one looks for factorized densities of the form:

$$q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau) = q(\boldsymbol{\Theta})\ q(\boldsymbol{\eta})\ q(\tau). \tag{3.18}$$

We make these expressions more specific in the next sections where we discuss the prior for $p_\Theta(\boldsymbol{\Theta})$ as well.

3.2.3 Prior specification for $\boldsymbol{\Theta}, \boldsymbol{\eta}, \boldsymbol{\mu}$ and $\boldsymbol{W}$

We discuss first the prior specification on $\boldsymbol{W}$. Its d_Θ columns $\boldsymbol{w}_i,\ i = 1, \ldots, d_\Theta$ span the subspace over which an approximation of $\boldsymbol{\Psi}$ is sought. We note that $\boldsymbol{\Psi}$ depends on the product $\boldsymbol{W}\boldsymbol{\Theta}$ which would remain invariant by appropriate rescaling of each pair

[1] The only additional requirement is that $\boldsymbol{H}$ is semi-positive definite or that a semi-positive approximation $\tilde{\boldsymbol{H}} \approx \boldsymbol{H}$ is used.

of $\boldsymbol{w}_i' = \alpha_i\ \boldsymbol{w}_i$ and $\Theta_i' = \frac{1}{\alpha_i}\Theta_i$ for any α_i. Hence, to resolve identifiability issues we require that $\boldsymbol{W}$ is *orthogonal*, i.e., $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_{d_\Theta}$ where $\boldsymbol{I}_{d_\Theta}$ is the $d_\Theta-$dimensional identity matrix. This is equivalent to employing a uniform prior on $\boldsymbol{W}$ on the Stiefel manifold $V_{d_\Theta}(\mathbb{R}^{d_\Psi})$ [182].

The latent, reduced coordinates $\boldsymbol{\Theta} \in \mathbb{R}^{d_\Theta}$ capture the variation of $\boldsymbol{\Psi}$ around its mean $\boldsymbol{\mu}$ along the directions of $\boldsymbol{W}$ as implied by Equation (3.4). Therefore, it is reasonable to assume that, a priori, these should have zero mean and should be uncorrelated [177]. For that purpose we adopt a multivariate Gaussian prior (denoted by $p_\Theta(\boldsymbol{\Theta})$ in the Equations of the previous section) with a diagonal covariance denoted by $\boldsymbol{\Lambda}_0^{-1} = diag(\lambda_{0,i}^{-1}), i = 1, \dots d_\Theta$

$$p_\Theta(\boldsymbol{\Theta}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_0^{-1}). \tag{3.19}$$

We select prior variances $\lambda_{0,i}^{-1}$ such that $\lambda_{0,1}^{-1} > \lambda_{0,2}^{-1} > \dots > \lambda_{0,d_\Theta}^{-1}$. This induces a natural (stochastic) ordering to the reduced coordinates $\boldsymbol{\Theta}$ since $\boldsymbol{\Psi}$ is invariant to permutations of the entries of the $\boldsymbol{\Theta}$ and the columns of $\boldsymbol{W}$ (Equation (3.4)). As a result of this ordering, Θ_1 is associated with the direction along which the largest variance in $\boldsymbol{\Psi}$ is attained, Θ_2 with the direction with the second largest variance and so on. We discuss the particular values given to prior hyperparameters $\lambda_{0,i}$ in the sequel (Section 3.3) and in Section 3.2.5 the possibility of an adaptive decomposition is also presented. This enables the sequential addition of reduced coordinates until a sufficiently good approximation to the posterior is attained.

As the role of the latent variables $\boldsymbol{\eta}$ is to capture any residual variance (that is not accounted for by $\boldsymbol{\Theta}$), we assume that, *a priori*, $\boldsymbol{\eta}$ can be modeled by a multivariate Gaussian that has zero mean and an isotropic covariance:

$$p_\eta(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{0}, \lambda_{0,\eta}^{-1}\boldsymbol{I}_{d_\Psi}). \tag{3.20}$$

The final aspect of the prior model pertains to $\boldsymbol{\mu}$: $p_\mu(\boldsymbol{\mu})$. We use a hierarchical prior that induces the requisite smoothness given that $\boldsymbol{\Psi}$ represents the spatial variability of the material parameters. In particular, the prior model employed penalizes the jumps in the values of Ψ_k and Ψ_l which correspond to neighboring sites/locations k, l. The definition of a neighborhood can be adjusted depending on the problem. In this work, we assume that sites/locations belong to the neighborhood if they correspond to adjacent pixels/voxels.[2] Suppose d_L is the total number of neighboring pairs of elements. Then for $m = 1, \dots, d_L$ and if k_m and l_m denote the corresponding neighboring pair:

$$p(\mu_{k_m} - \mu_{l_m}|\xi_m) = \sqrt{\frac{\xi_m}{2\pi}} e^{-\frac{\xi_m}{2}(\mu_{k_m} - \mu_{l_m})^2}. \tag{3.21}$$

[2]This results in four neighbors for a single quadrilateral element and in three neighbors per element for triangular elements (less for elements at the boundary).

The strength of the penalty is proportional to the hyperparameter $\xi_m > 0$, i.e., smaller values of ξ_m induce a weaker penalty and vice versa [183]. Let $\boldsymbol{L}$ the $d_L \times d_\Psi$ denote the Boolean matrix that can be used to produce the vector of all d_L jumps (as the one above) between all neighboring sites from the vector $\boldsymbol{\mu}$ as $\boldsymbol{L\mu}$.[3] $\boldsymbol{\Xi} = diag(\xi_m)$ is the *diagonal matrix* containing all the hyperparameters ξ_j associated with each of these jumps. We can represent the combined prior on $\boldsymbol{\mu}$ as:

$$p(\boldsymbol{\mu}|\boldsymbol{\Xi}) \propto |\boldsymbol{\Xi}|^{1/2} e^{-\frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{L}^T \boldsymbol{\Xi} \boldsymbol{L} \boldsymbol{\mu}}. \tag{3.22}$$

A conjugate prior of the hyperparameters $\boldsymbol{\Xi}$ is a product of Gamma distributions:

$$p_\Xi(\boldsymbol{\Xi}) = \prod_{m=1}^{d_L} Gamma(a_\xi, b_\xi). \tag{3.23}$$

As in [183], the independence is motivated by the absence of correlation (a priori) with regard to the locations of the jumps. In this work we use $a_\xi = b_\xi = 0$ which corresponds to a limiting case of a Jeffreys prior that is scale invariant. We note that in contrast to previous works where such priors have been employed for the vector of unknowns $\boldsymbol{\Psi}$ and MAP estimates have been obtained [131], we employ this here for $\boldsymbol{\mu}$ which is only part of the overall decomposition in Equation (3.4). We discuss in the following section the update equations for $\boldsymbol{\mu}$ and the associated hyper-parameters $\boldsymbol{\Xi}$ as well as for the remaining model variables. Furthermore, an Expectation-Maximization scheme to derive $\log\, p_\mu(\boldsymbol{\mu})$ will be derived (cf. also Appendix A).

3.2.4 Update equations for $q(\boldsymbol{\Theta}), q(\boldsymbol{\eta}), q(\tau), \boldsymbol{\mu}, \boldsymbol{W}$

We postulate that the reduced coordinates $\boldsymbol{\Theta}$ as well as $\boldsymbol{\eta}$ should, a posteriori, have zero mean as they capture variability around $\boldsymbol{\mu}$ and the residual "noise" respectively. For that purpose we confine our search for $q(\boldsymbol{\Theta})$, $q(\boldsymbol{\eta})$ to distributions with zero mean. Given the aforementioned priors and the linearization discussed in the previous section, we can readily deduce from the lower bound, Equation (3.13), that the optimal approximate posteriors $q^{opt}(\boldsymbol{\Theta})$, $q^{opt}(\boldsymbol{\eta})$ and $q^{opt}(\tau)$, under the mean-field Variational Bayesian scheme adopted, will be:

$$\begin{aligned} q^{opt}(\boldsymbol{\Theta}) &\equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1}), \\ q^{opt}(\boldsymbol{\eta}) &\equiv \mathcal{N}(\mathbf{0}, \lambda_\eta^{-1} \boldsymbol{I}_{d_\Psi}), \\ q^{opt}(\tau) &\equiv Gamma(a, b). \end{aligned} \tag{3.24}$$

The associated parameters are given by the following *iterative* equations:

$$\begin{aligned} a &= a_0 + d_y/2, \\ b = b_0 + \tfrac{1}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 &+ \tfrac{1}{2} tr(\boldsymbol{W}^T \boldsymbol{G}^T \boldsymbol{G} \boldsymbol{W} \boldsymbol{\Lambda}^{-1}) + \tfrac{1}{2} \lambda_\eta^{-1} tr(\boldsymbol{G}^T \boldsymbol{G}), \end{aligned} \tag{3.25}$$

[3] $\boldsymbol{L}$ has zero entries which are adjusted for $m = 1 : d_L$ by $\boldsymbol{L}(m, k_m) = \boldsymbol{L}(m, k_m) + 1$ and $\boldsymbol{L}(m, l_m) = \boldsymbol{L}(m, l_m) - 1$.

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 + < \tau >_\tau \boldsymbol{W}^T \boldsymbol{G}^T \boldsymbol{G} \boldsymbol{W}, \tag{3.26}$$

$$\lambda_\eta = \lambda_{0,\eta} + \frac{1}{d_\Psi} < \tau >_\tau tr(\boldsymbol{G}^T \boldsymbol{G}), \tag{3.27}$$

where $< \tau >=< \tau >_\tau =< \tau >_q = \frac{a}{b}$.

As a result of the aforementioned equations and Equation (3.4), one can establish that the *posterior* of $\boldsymbol{\Psi}$ is approximated by a Gaussian with mean and covariance given by:

$$\begin{array}{l} < \boldsymbol{\Psi} >_q =< \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} >_q = \boldsymbol{\mu}, \\ \mathrm{Cov}[\boldsymbol{\Psi}] = \boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T + \lambda_\eta^{-1}\boldsymbol{I}_{d_\Psi}. \end{array} \tag{3.28}$$

We note that if we diagonalize $\boldsymbol{\Lambda}^{-1}$, i.e., $\boldsymbol{\Lambda}^{-1} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$ where $\boldsymbol{D}$ is diagonal and $\boldsymbol{V}$ is orthogonal with columns equal to the eigenvectors of $\boldsymbol{\Lambda}^{-1}$, then:

$$\begin{array}{rl} \mathrm{Cov}[\boldsymbol{\Psi}] & = \boldsymbol{W}\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{W}^T + \lambda_\eta^{-1}\boldsymbol{I}_{d_\Psi} \\ & = \tilde{\boldsymbol{W}}\boldsymbol{D}\tilde{\boldsymbol{W}}^T + \lambda_\eta^{-1}\boldsymbol{I}_{d_\Psi}. \end{array} \tag{3.29}$$

$\tilde{\boldsymbol{W}}$ is also orthogonal (i.e., $\tilde{\boldsymbol{W}}^T\tilde{\boldsymbol{W}} = \boldsymbol{I}_{d_\Theta}$) and contains the d_Θ principal directions of the posterior covariance of $\boldsymbol{\Psi}$. Hence, it suffices to consider approximate posteriors $q(\boldsymbol{\Theta})$ with covariance $\boldsymbol{\Lambda}^{-1}$ that is *diagonal*, i.e., $\boldsymbol{\Lambda} = diag(\lambda_i),\ i = 1, \ldots, d_\Theta$. In this case the update equations for λ_i in Equation (3.26) reduce to:

$$\lambda_i = \lambda_{0,i} + < \tau >_\tau \boldsymbol{w}_i^T \boldsymbol{G}^T \boldsymbol{G} \boldsymbol{w}_i. \tag{3.30}$$

We would like to point out that, despite the prior assumption on uncorrelated $\boldsymbol{\Theta}$, the posterior on $\boldsymbol{\Psi}$ exhibits correlation and captures the principal directions along which the variance is largest. Furthermore, implicit to the aforementioned derivations is the assumption of a unimodal posterior on $\boldsymbol{\Theta}$ and subsequently on $\boldsymbol{\Psi}$. This assumption can be relaxed by employing a mixture of Gaussians (e.g., [93]) that will enable the approximation of highly non-Gaussian and potentially multimodal posteriors. Such approximations could also be combined with the employment of different basis sets $\boldsymbol{W}$ for each of the mixture component which would provide a wide range of possibilities. We defer further discussions along these lines to Section 4. In the examined elastography applications, the unimodal assumption seems to be a reasonable, due to generally large amounts of data/observations obtained from various imaging modalities.

Given the aforementioned results one can obtain an expression for the variational

lower bound $\mathcal{F}$ in Equation (3.13):

$$\begin{aligned}\mathcal{F}(q(\boldsymbol{\Theta}), q(\boldsymbol{\eta}), q(\tau), \boldsymbol{\mu}, \boldsymbol{W}) &= \left\langle \log \frac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta},\boldsymbol{\eta},\tau,\boldsymbol{\mu},\boldsymbol{W})\ p_{\Theta}(\boldsymbol{\Theta})\ p_{\eta}(\boldsymbol{\eta})\ p_{\tau}(\tau)}{q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau)} \right\rangle_q \\ &+ \log p_{\mu}(\boldsymbol{\mu}) + \log p_W(\boldsymbol{W}) \\ &= -\tfrac{d_y}{2}\log 2\pi + \tfrac{d_y}{2} < \log\tau >_{\tau} \\ &-\tfrac{<\tau>}{2} < ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}) - \boldsymbol{G}(\boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta})||^2 >_q \\ &+\tfrac{1}{2}\log|\boldsymbol{\Lambda}_0| - \tfrac{1}{2}\boldsymbol{\Lambda}_0 :< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}} + \tfrac{d_{\Psi}}{2}\log\lambda_{0,\eta} \\ &-\tfrac{\lambda_{0,\eta}}{2}\boldsymbol{I} :< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{\boldsymbol{\eta}} \\ &+(a_0 - 1) < \log\tau >_{\tau} - b_0 < \tau >_{\tau} - \log Z(a_0, b_0) \\ &-\tfrac{1}{2}\log|\boldsymbol{\Lambda}| + \tfrac{d_{\Theta}}{2} - \tfrac{d_{\Psi}}{2}\log\lambda_{\eta} + \tfrac{d_{\Psi}}{2} \\ &-(a-1) < \log\tau >_{\tau} + b < \tau >_{\tau} + \log Z(a, b) \\ &+ \log p_{\mu}(\boldsymbol{\mu}) + \log p_W(\boldsymbol{W}),\end{aligned} \tag{3.31}$$

where $Z(\gamma, \delta) = \frac{\Gamma(\gamma)}{\delta^{\gamma}}$ is the normalization constant of a $Gamma$ distribution with parameters γ, δ. The aforementioned equation can be further simplified by making use of the following expectations: $< \boldsymbol{\Theta} >_{\boldsymbol{\Theta}} = \mathbf{0}$, $< \boldsymbol{\eta} >_{\boldsymbol{\eta}} = \mathbf{0}$, $< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}} = \boldsymbol{\Lambda}^{-1}$, $< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{\boldsymbol{\eta}} = \lambda_{\eta}^{-1}\boldsymbol{I}_{d_{\Psi}}$:

$$\begin{aligned}\mathcal{F}(q^{opt}(\boldsymbol{\Theta}), q^{opt}(\boldsymbol{\eta}), q^{opt}(\tau), \boldsymbol{\mu}, \boldsymbol{W}) &= -\tfrac{d_y}{2}\log 2\pi + \tfrac{d_y}{2} < \log\tau >_{\tau} \\ &-\tfrac{<\tau>}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 \\ &-\tfrac{<\tau>}{2}\boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\ \boldsymbol{W} : \boldsymbol{\Lambda}^{-1} \\ &-\tfrac{<\tau>}{2}\boldsymbol{G}^T\boldsymbol{G} : \lambda_{\eta}^{-1}\boldsymbol{I}_{d_{\Psi}} \\ &+\tfrac{1}{2}\log|\boldsymbol{\Lambda}_0| - \tfrac{1}{2}\boldsymbol{\Lambda}_0 : \boldsymbol{\Lambda}^{-1} + \tfrac{d_{\Psi}}{2}\log\lambda_{0,\eta} \\ &-\tfrac{d_{\Psi}\lambda_{0,\eta}}{2\lambda_{\eta}} + (a_0 - 1) < \log\tau >_{\tau} - b_0 < \tau >_{\tau} \\ &- \log Z(a_0, b_0) \\ &-\tfrac{1}{2}\log|\boldsymbol{\Lambda}| + \tfrac{d_{\Theta}}{2} - \tfrac{d_{\Psi}}{2}\log\lambda_{\eta} \\ &+\tfrac{d_{\Psi}}{2} - (a-1) < \log\tau >_{\tau} + b < \tau >_{\tau} \\ &+ \log Z(a, b) + \log p_{\mu}(\boldsymbol{\mu}) + \log p_W(\boldsymbol{W}).\end{aligned} \tag{3.32}$$

In order to update $\boldsymbol{W}$ in the VB-Maximization step, it suffices to consider only the terms of $\mathcal{F}$ that depend on it which we denote by $\mathcal{F}_W(\boldsymbol{W})$, i.e.:

$$\mathcal{F}_W(\boldsymbol{W}) = -\tfrac{<\tau>_{\tau}}{2}\boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\ \boldsymbol{W} : \boldsymbol{\Lambda}^{-1} + \log p_W(\boldsymbol{W}). \tag{3.33}$$

As discussed earlier the prior $p_W(\boldsymbol{W})$ enforces the orthogonality constraint on $\boldsymbol{W}$. To address this constrained optimization problem, we employ the iterative algorithm proposed in [184] which has proven highly efficient in terms of the number of iterations and the cost per iterations in several settings. It employs the Cayley transform [185] to preserve the constraint during the optimization and makes use only of first order derivatives:

$$\frac{\partial\mathcal{F}_W}{\partial\boldsymbol{W}} = - < \tau >_{\tau} \boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{\Lambda}^{-1} + \frac{\partial \log p_W(\boldsymbol{W})}{\partial\boldsymbol{W}}, \tag{3.34}$$

with

$$\boldsymbol{B} = \frac{\partial \mathcal{F}_W}{\partial \boldsymbol{W}} \boldsymbol{W}^T - \boldsymbol{W} \frac{\partial \mathcal{F}_W}{\partial \boldsymbol{W}}^T. \tag{3.35}$$

The update equations are based on a Crank-Nicholson-like scheme:

$$\boldsymbol{W}_{new} = (\boldsymbol{I}_{d_\Psi} + \frac{\alpha_W}{2}\boldsymbol{B})^{-1}(\boldsymbol{I}_{d_\Psi} + \frac{\alpha_W}{2}\boldsymbol{B})\boldsymbol{W}_{old}, \tag{3.36}$$

where α_W is the step size and $\boldsymbol{I}_{d_\Psi} \in \mathbb{R}^{d_\Psi \times d_\Psi}$ the identity matrix with the dimension of $\boldsymbol{\Psi}$, d_Ψ. One notes that the aforementioned update preserves the orthogonality of $\boldsymbol{W}_{new}$ ([184]). In order to derive a good step size we use the Barzilai-Borwein scheme [186] which results in a non-monotone line search algorithm:

$$\alpha_W = \frac{||tr(\Delta \boldsymbol{W} \Delta \frac{\partial \mathcal{F}_W}{\partial \boldsymbol{W}})||}{tr(\Delta \frac{\partial \mathcal{F}_W}{\partial \boldsymbol{W}}^T \Delta \frac{\partial \mathcal{F}_W}{\partial \boldsymbol{W}})}, \tag{3.37}$$

where Δ represents the difference between the current parameter values as compared to the previous step and the absolute value of the denominator is taken such that α_W is never negative. As discussed in detail in [184] the inversion of the $d_\Psi \times d_\Psi$ matrix $(\boldsymbol{I}_{d_\Psi} + \frac{\alpha_W}{2}\boldsymbol{B})$ in Equation (3.36) can be efficiently performed by inverting a matrix of dimension $2d_\Theta$ which is much smaller than d_Ψ. We remark that the updates of $\boldsymbol{W}$ require no forward calls for the computation of $\boldsymbol{y}(\boldsymbol{\mu})$ or its derivatives $\boldsymbol{G}$. The updates/iterations are terminated when no further improvement to the objective is possible.

The final component involves the optimization of $\boldsymbol{\mu}$. As with $\boldsymbol{W}$ we consider only the terms of $\mathcal{F}$ (Equation (3.32)) that depend on $\boldsymbol{\mu}$ which we denote by $\mathcal{F}_\mu(\boldsymbol{\mu})$, i.e.:

$$\mathcal{F}_\mu(\boldsymbol{\mu}) \quad = -\tfrac{<\tau>_\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 + \log p_\mu(\boldsymbol{\mu}). \tag{3.38}$$

Due to the analytical unavailability of $\log p_\mu(\boldsymbol{\mu})$ and its derivatives $\frac{\partial \log p_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$, we employ here an Expectation-Maximization scheme [187, 135] which we describe in Appendix A for completeness. The output of this algorithm is also the posterior on the hyperparameters ξ_m, Equation (3.21), which captures the locations of jumps in $\boldsymbol{\mu}$ as well as the probabilities associated with them. The cost of the numerical operations is minimal and scales linearly with the number of neighboring pairs d_L. In the following, we simply make use of Equations (A.3) without further explanation.

Formally, the determination of the optimal $\boldsymbol{\mu}$ would require the derivatives $\frac{\partial \mathcal{F}_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ in Equation (3.38). We note that $\boldsymbol{G} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}|_{\boldsymbol{\Psi}=\boldsymbol{\mu}}$ depends on $\boldsymbol{\mu}$. Hence, finding $\frac{\partial \mathcal{F}_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ would require the computation of second-order derivatives of $\boldsymbol{y}(\boldsymbol{\Psi})$ which poses significantly computational difficulties in the high-dimensional considered setting. To avoid this and *only* for the purpose of the $\boldsymbol{\mu}$-updates, we linearize Equation (3.38) around the current guess by ignoring the dependence of $\boldsymbol{G}$ on $\boldsymbol{\mu}$ or equivalently by

assuming that $\boldsymbol{G}$ remains constant in the vicinity of the current guess. In particular, let $\boldsymbol{\mu}^{(t)}$ denote the value of $\boldsymbol{\mu}$ at iteration t, then in order to find the increment $\Delta\boldsymbol{\mu}^{(t)}$, we define the new objective $F_{\mu}^{(t)}(\Delta\boldsymbol{\mu}^{(t)})$ as follows:

$$\begin{aligned} F_{\mu}^{(t)}(\Delta\boldsymbol{\mu}^{(t)}) = \; & F_{\mu}(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)}) + \log p(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)}) \\ & = -\tfrac{<\tau>_{\tau}}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)})||^2 \\ & -\tfrac{1}{2}(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)})^T\boldsymbol{L}^T < \boldsymbol{\Xi} >_{\boldsymbol{\Xi}} \boldsymbol{L}(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)}) \\ & \approx -\tfrac{<\tau>_{\tau}}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}^{(t)}) - \boldsymbol{G}^{(t)}\Delta\boldsymbol{\mu}^{(t)}||^2 \\ & -\tfrac{1}{2}(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)})^T\boldsymbol{L}^T < \boldsymbol{\Xi} >_{\boldsymbol{\Xi}} \boldsymbol{L}(\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)}). \end{aligned} \tag{3.39}$$

We remark that there is no approximation with regard to the $p_{\mu}(\boldsymbol{\mu})$ prior term. By keeping only the terms depending on $\Delta\boldsymbol{\mu}^{(t)}$ in the equation above we obtain:

$$\begin{aligned} F_{\mu}^{(t)}(\Delta\boldsymbol{\mu}^{(t)}) \; & = -\tfrac{<\tau>_{\tau}}{2}(\Delta\boldsymbol{\mu}^{(t)})^T(\boldsymbol{G}^{(t)})^T\boldsymbol{G}^{(t)}\;\Delta\boldsymbol{\mu}^{(t)} \\ & + <\tau>_{\tau}\,(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}^{(t)}))^T\boldsymbol{G}^{(t)}\;\Delta\boldsymbol{\mu}^{(t)} \\ & -\tfrac{1}{2}(\Delta\boldsymbol{\mu}^{(t)})^T\boldsymbol{L}^T < \boldsymbol{\Xi} >_{\boldsymbol{\Xi}} \boldsymbol{L}\;\Delta\boldsymbol{\mu}^{(t)} \\ & -(\boldsymbol{\mu}^{(t)})^T\boldsymbol{L}^T < \boldsymbol{\Xi} >_{\boldsymbol{\Xi}} \boldsymbol{L}\;\Delta\boldsymbol{\mu}^{(t)}. \end{aligned} \tag{3.40}$$

This is concave and quadratic with respect to the unknown $\Delta\boldsymbol{\mu}^{(t)}$. The maximum can be found by setting $\frac{\partial F_{\mu}^{(t)}(\Delta\boldsymbol{\mu}^{(t)})}{\partial\Delta\boldsymbol{\mu}^{(t)}} = \boldsymbol{0}$, which yields:

$$\begin{aligned} & (<\tau>_{\tau}\,(\boldsymbol{G}^{(t)})^T\boldsymbol{G}^{(t)} + \boldsymbol{L}^T < \boldsymbol{\Xi} >_{\boldsymbol{\Xi}} \boldsymbol{L})\Delta\boldsymbol{\mu}^{(t)} \\ & =<\tau>_{\tau}\,(\boldsymbol{G}^{(t)})^T(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}^{(t)})) - \boldsymbol{L}^T < \boldsymbol{\Xi} >_{\boldsymbol{\Xi}} \boldsymbol{L}\boldsymbol{\mu}^{(t)}. \end{aligned} \tag{3.41}$$

We note that the *exact* objective $F_{\mu}(\boldsymbol{\mu}) + \log p_{\mu}(\boldsymbol{\mu})$ is evaluated at $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\mu}^{(t+1)}$ is accepted only if the value of the objective is larger than that at $\boldsymbol{\mu}^{(t)}$. Iterations are terminated when no further improvement is possible. Finally, it was found that activating the regularization term ($\log p_{\mu}(\boldsymbol{\mu})$) after five updates/iterations during which the optimization is performed solely on the basis of $F_{\mu}(\boldsymbol{\mu})$, enabled better exploration of the feasible solutions. This addresses first an optimization w.r.t. $F_{\mu}(\boldsymbol{\mu})$ before also the smoothing prior is incorporated.

We summarize below the basic steps of the iterative Variational Bayesian scheme proposed in Algorithm 1.

Algorithm 1 Variational Bayesian Approach Including Dictionary Learning for fixed d_{Θ}

1: Initialize $\boldsymbol{\mu}$, $\boldsymbol{W}$, $\boldsymbol{\Lambda}_0$, $\lambda_{0,\eta}$ and the hyperparameters a_0, b_0, a_{ξ}, b_{ξ}
2: Update $\boldsymbol{\mu}$ using Equation (3.41)
3: **while** $\mathcal{F}$ (Equation (3.32)) has not converged **do**
4: Update $\boldsymbol{W}$ using Equations (3.33-3.37)
5: Update $q(\boldsymbol{\Theta}) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1})$, $q(\boldsymbol{\eta}) \equiv \mathcal{N}(\boldsymbol{0}, \lambda_{\eta}^{-1}\boldsymbol{I}_{d_{\Psi}})$ using Equation (3.30), Equation (3.27) and $q(\tau) \equiv Gamma(a, b)$ using Equation (3.25)
6: **end while**

With regard to the overall computational cost we note that the updates of $\boldsymbol{\mu}$ are the most demanding as they require calls to the forward model to evaluate $\boldsymbol{y}(\boldsymbol{\mu}^{(t)})$ and the derivatives $\boldsymbol{G}^{(t)} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}|_{\boldsymbol{\Psi}=\boldsymbol{\mu}^{(t)}}$, as described in Appendix D. The updates were terminated when no further increase in $\mathcal{F}$ (Equation (3.32)) can be attained.

3.2.5 Adaptive learning - Cardinality of reduced coordinates

The presentation thus far was based on a fixed number d_Θ of reduced coordinates $\boldsymbol{\Theta}$. A natural question that arises is how many should one consider. In order to address this issue, we propose an adaptive learning scheme. According to this, the analysis is first performed with a few (even one) reduced coordinates and upon convergence additional reduced coordinates are introduced, either in small batches or even one-by-one. Critical to the implementation of such a scheme is a metric for the progress achieved by the addition of reduced coordinates and basis vectors which can also be used as a termination criterion.

In this work, we advocate the use of an information-theoretic criterion which measures the information gain between the prior beliefs on $\boldsymbol{\Theta}$ and the corresponding posterior. To measure such gains, we employ again the KL-divergence between the aforementioned distributions. In particular, if $p_{d_\Theta}(\boldsymbol{\Theta})$ (Section 3.2.3) and $q_{d_\Theta}(\boldsymbol{\Theta})$ (Equation (3.30)) denote the d_Θ-dimensional prior and posterior respectively, we define the quantity $I(d_\Theta)$ as follows:

$$I(d_\Theta) = \frac{KL(p_{d_\Theta}(\boldsymbol{\Theta})||q_{d_\Theta}(\boldsymbol{\Theta})) - KL(p_{d_\Theta-1}(\boldsymbol{\Theta})||q_{d_\Theta-1}(\boldsymbol{\Theta}))}{KL(p_{d_\Theta}(\boldsymbol{\Theta})||q_{d_\Theta}(\boldsymbol{\Theta}))}, \tag{3.42}$$

which measures the (relative) information gain from $d_\Theta - 1$ to d_Θ reduced coordinates. The KL divergence between $p_{d_\Theta}(\boldsymbol{\Theta})$ and $q_{d_\Theta}(\boldsymbol{\Theta})$, with $p_{d_\Theta}(\boldsymbol{\Theta}) \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_0^{-1})$ and $q_{d_\Theta}(\boldsymbol{\Theta}) \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1})$ where $\boldsymbol{\Lambda_0}, \boldsymbol{\Lambda}$ are diagonal as explained previously, follows with:

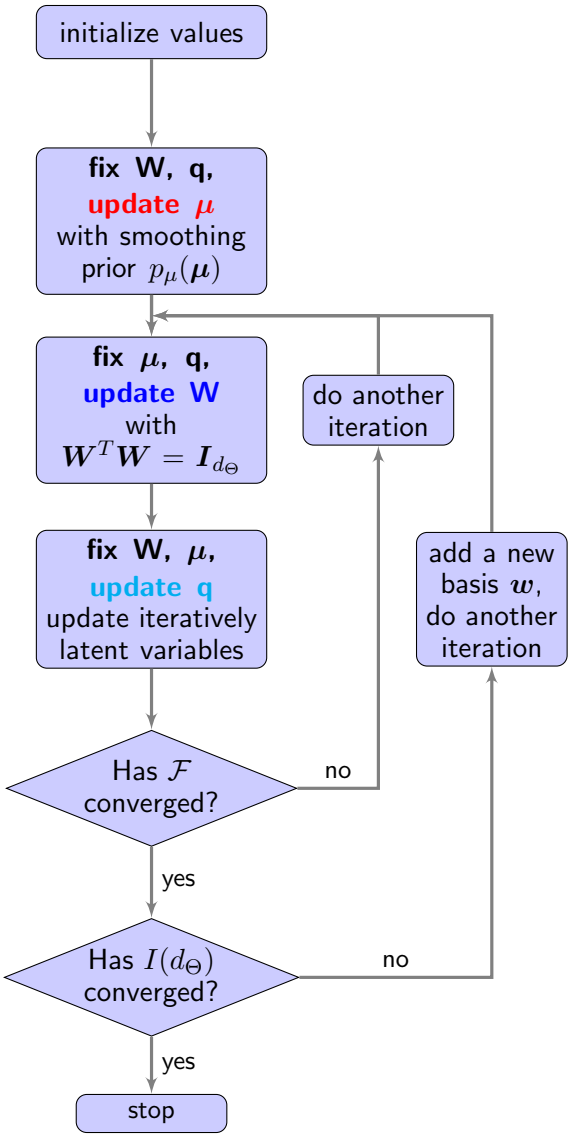
$$KL(p_{d_\Theta}(\boldsymbol{\Theta})||q_{d_\Theta}(\boldsymbol{\Theta})) = \frac{1}{2}\sum_{i=1}^{d_\Theta}(-\log(\frac{\lambda_i}{\lambda_{0,i}}) + \frac{\lambda_i}{\lambda_{0,i}} - 1), \tag{3.43}$$

and Equation (3.42) becomes:

$$I(d_\Theta) = \frac{\sum_{i=1}^{d_\Theta}(-\log(\frac{\lambda_i}{\lambda_{0,i}}) + \frac{\lambda_i}{\lambda_{0,i}} - 1) - \sum_{i=1}^{d_\Theta-1}(-\log(\frac{\lambda_i}{\lambda_{0,i}}) + \frac{\lambda_i}{\lambda_{0,i}} - 1)}{\sum_{i=1}^{d_\Theta}(-\log(\frac{\lambda_i}{\lambda_{0,i}}) + \frac{\lambda_i}{\lambda_{0,i}} - 1)}. \tag{3.44}$$

In the simulations performed in Section 3.3, we demonstrate the evolution of this metric as reduced-coordinates/basis vectors are added one-by-one. The addition of reduced coordinates was terminated when $I(d_\Theta)$ was below 1% for at least five consecutive d_Θ. In Figure 3.3, an overview flowchart of the proposed algorithm is shown.

It incorporates the VB algorithm including dictionary learning from Algorithm 1 and the information gain assessment to identify the necessary number of basis vectors from this subsection.



μ-update:

$$arg\max_{\boldsymbol{\mu}} \mathcal{F}_\mu = -\frac{<\tau>_\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 + \log\, p_\mu(\boldsymbol{\mu})$$

W-update:

$$arg\max_{\boldsymbol{W}} \mathcal{F}_W = -\frac{<\tau>_\tau}{2}\boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W} : \boldsymbol{\Lambda}^{-1} + \log\, p_W(\boldsymbol{W})$$

q-update:

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 + <\tau>_\tau \boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W}$$

$$\lambda_\eta = \lambda_{0,\eta} + \frac{1}{d_\Psi} <\tau>_\tau tr(\boldsymbol{G}^T\boldsymbol{G})$$

$$a = a_0 + d_y/2$$

$$b = b_0 + \frac{1}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu})||^2 + \frac{1}{2}tr(\boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}) + \frac{1}{2}\lambda_\eta^{-1}tr(\boldsymbol{G}^T\boldsymbol{G})$$

Figure 3.3: Flowchart for the new algorithm. As the $\boldsymbol{\mu}$-update does not depend on $\boldsymbol{W}$ just one $\boldsymbol{\mu}$-update (which is the expensive part of the full algorithm) is necessary during the calculations.

3.2.6 Validation- Combining VB approximations with importance sampling

Thus far we have employed the variational lower bound in order to identify the optimal dimensionality reduction and to infer the latent variables that approximate the posterior. The goal of this section is twofold. Firstly, to show how the biased VB approximation can be used in order to obtain efficiently, *(asymptotically) unbiased* estimates with regard to the true posterior and secondly, to assess (quantitatively) the accuracy of the VB approximation. To that end, we employ importance sampling (Section 2.2.3) with the variational posterior as the importance sampling distribution. We can thus obtain consistent estimators of several exact posterior quantities as well as of measure the efficiency of importance sampling (IS).

The performance of IS can decay rapidly in high dimensions [79] and due to the fact that $\boldsymbol{\eta}$ has a negligible effect in the inferred posterior (as seen in the discussed examples), we propose using the *exact* posterior $p(\boldsymbol{\Theta}|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W}) = \frac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta},\boldsymbol{\mu},\boldsymbol{W})\ p_\Theta(\boldsymbol{\Theta})}{p(\hat{\boldsymbol{y}}|\boldsymbol{\mu},\boldsymbol{W})}$ as the target density. We note that when τ is unknown, as in the cases considered herein, the (marginal) likelihood $p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{W})$ can be determined by integrating with respect to τ. With the conjugate Gamma prior adopted (Equation (3.3)) this can be done analytically and would yield:

$$\begin{aligned} p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{W}) &= \int p(\hat{\boldsymbol{y}}, \tau|\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{W})\ d\tau \\ &= \int p(\hat{\boldsymbol{y}}|\tau, \boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{W})\ p_\tau(\tau)\ d\tau \\ &\propto \frac{\Gamma(a_0+d_y/2)}{(b_0+\frac{||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\mu}+\boldsymbol{W}\boldsymbol{\Theta})||^2}{2})^{a_0+d_y/2}}. \end{aligned} \tag{3.45}$$

In cases where non-conjugate priors for τ are employed, the IS procedure detailed here has to be performed in the joint space $(\boldsymbol{\Theta}, \tau)$.

The evidence is:

$$p(\hat{\boldsymbol{y}}|\boldsymbol{\mu}, \boldsymbol{W}) = \int p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{W})\ p_\Theta(\boldsymbol{\Theta})\ d\boldsymbol{\Theta}, \tag{3.46}$$

and the expectation of any function $g(\boldsymbol{\Psi}) = g(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta})$ with regard to the *exact posterior* $p(\boldsymbol{\Theta}|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W})$:

$$\begin{aligned} < g(\boldsymbol{\Psi}) >_{p(\boldsymbol{\Theta}|\hat{\boldsymbol{y}},\boldsymbol{\mu},\boldsymbol{W})} &= \int g(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta})\ p(\boldsymbol{\Theta}|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W})\ d\boldsymbol{\Theta} \\ &= \int g(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta})\ \frac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta},\boldsymbol{\mu},\boldsymbol{W})\ p_\Theta(\boldsymbol{\Theta})}{p(\hat{\boldsymbol{y}}|\boldsymbol{\mu},\boldsymbol{W})}\ d\boldsymbol{\Theta}, \end{aligned} \tag{3.47}$$

can be estimated using IS with respect to the IS density $q(\boldsymbol{\Theta})$ as follows:

$$\begin{aligned} &\frac{1}{M}\sum_{m=1}^{M} w(\boldsymbol{\Theta}^{(m)}) \ \rightarrow\ p(\hat{\boldsymbol{y}}|\boldsymbol{\mu}, \boldsymbol{W}), \\ &\frac{1}{\sum_{m=1}^{M} w(\boldsymbol{\Theta}^{(m)})}\sum_{m=1}^{M} g(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta}^{(m)})\ w(\boldsymbol{\Theta}^{(m)}) \ \rightarrow\ \ < g(\boldsymbol{\Psi}) >_{p(\boldsymbol{\Theta}|\hat{\boldsymbol{y}},\boldsymbol{\mu},\boldsymbol{W})}. \end{aligned} \tag{3.48}$$

The samples $\{\boldsymbol{\Theta}^{(m)}\}_{m=1}^{M}$ are independent draws from $q(\boldsymbol{\Theta})$ and the IS weights are given by:

$$w(\boldsymbol{\Theta}) = \frac{p(\hat{\boldsymbol{y}}|\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{W})\ p_{\Theta}(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta})}. \tag{3.49}$$

The (normalized) effective sample size from Equation (2.11) to measure the degeneracy in the population of particles/samples as quantified by their variance [188] is:

$$ESS_{IS} = \frac{(\sum_{m=1}^{M} w(\boldsymbol{\Theta}^{(m)}))^2}{M \sum_{m=1}^{M} w^2(\boldsymbol{\Theta}^{(m)})}. \tag{3.50}$$

In summary, the VB framework advocated introduces approximations due to the linearization of the response (Equation (3.15)) and the mean field approximation (Equation (3.18)). To assess the bias of these approximations in the posterior inferred, we employ IS as explained above. This can lead to accuracy metrics (e.g., ESS) but more importantly can produce (asymptotically) unbiased statistics with regard to the exact posterior, i.e., the one obtained without the approximations mentioned earlier. These metrics can be readily compared with those of alternative strategies (e.g., MCMC as in Equation (2.15)). Unequivocally, another important source of error is due to model discrepancies. That is, if the difference between observables and model predictions in Equation (3.1) is not valid due missing physics, discretization errors, etc., then the inference results will deviate from reality, irrespectively of the numerical tools one employs [189, 190, 108]. We emphasize that the methodology proposed, as most strategies for the solution of inverse problems, is based on the assumption that model errors are zero or in any case much smaller than the observation errors.

3.3 Numerical illustration

The elastography examples presented are concerned with the probabilistic identification of unknown material parameters from interior measured displacement data. We demonstrate the efficacy of the proposed methodology in two, two-dimensional cases where synthetic displacement data are utilized. The data are contaminated with noise as discussed below. The first example is based on a linear elastic material model. The second example incorporates the Mooney-Rivlin material model which is used to model nonlinear and incompressible response.

In the computations, we use $a_0 = b_0 = a_\xi = b_\xi = 0$. We employ the adaptive learning scheme discussed in Section 3.2.5 whereby reduced-coordinates/basis vectors are added one-by-one. The first reduced coordinate is assigned the broadest prior, i.e., $\lambda_{0,1}$ is the smallest of all other $\lambda_{0,i}$ and captures the largest expected (a priori) variance. For subsequent bases $i = 2, 3, \ldots$ we assign values to the precision parameters $\lambda_{0,i}$ as follows:

$$\lambda_{0,i} = max(\lambda_{0,1}, \lambda_{i-1} - \lambda_{0,i-1}), \quad i = 2, 3, \ldots, d_\Theta, \tag{3.51}$$

We note that λ_{i-1} corresponds to the *posterior* precision for the *previous* reduced coordinate Θ_{i-1} as found in Equation (3.30) according to which $\lambda_{0,i} =< \tau >_\tau \boldsymbol{w}_{i-1}^T \boldsymbol{G}^T \boldsymbol{G} \boldsymbol{w}_{i-1}$. This essentially implies that, a priori, the next reduced coordinate Θ_i will have the precision of the previous one as long as it is larger than the threshold $\lambda_{0,1}$. Since by construction $\boldsymbol{w}_i^T \boldsymbol{G}^T \boldsymbol{G} \boldsymbol{w}_i > \boldsymbol{w}_{i-1}^T \boldsymbol{G}^T \boldsymbol{G} \boldsymbol{w}_{i-1}$, we have that $\lambda_{0,i+1} \geq \lambda_{0,i}$. For the prior of $\boldsymbol{\eta}$ we use $\lambda_{0,\eta} = \max_i(\lambda_{0,i})$ as $\boldsymbol{\eta}$ represents the residual variance which is a priori smaller than the smallest variance of the reduced coordinates $\boldsymbol{\Theta}$.

The most important quantities and dimensions of the ensuing two examples are summarized in Table 3.1.

	Example 1	Example 2
Dimension of observables: d_y	198	5100
Dimension of latent variables: d_Ψ	90	2500
Dimension of reduced latent variables: d_Θ	$5-10$	$15-25$
Nr. of forward calls	< 25	< 35

Table 3.1: Summary of the number of observables, forward calls and the dimensionality reduction in the following two examples.

Details about the implemented software can be found in Appendix E.

Example 1: Linear elastic material

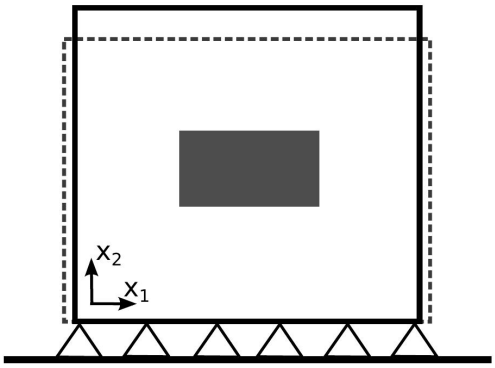
The primary objective of the first example is to assess the performance of the proposed framework in terms of accuracy and dimensionality reduction in a simple problem with the absence of model errors. For that purpose, we consider a linear, isotropic elastic material model where the stress-strain relation is given by:

$$\boldsymbol{S} = \mathbb{C} : \boldsymbol{E}, \tag{3.52}$$

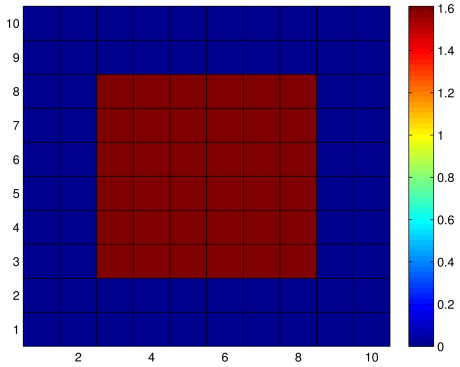
where $\mathbb{C}$ is the elasticity tensor [141]. It is given by:

$$\mathbb{C} = \frac{E}{(1+\nu)}(\boldsymbol{I} + \frac{\nu}{(1-2\nu)}\mathbf{1} \otimes \mathbf{1}), \tag{3.53}$$

and E is the elastic modulus. The second material parameter is Poisson's ratio ν which in this example is assumed to be known ($\nu = 0$). The vector of unknown parameters $\boldsymbol{\Psi}$ consists of the values of the elastic moduli at each finite element. We assume that the elastic modulus can take two values $E_{inclusion}$ and E_{matrix} such that $\frac{E_{inclusion}}{E_{matrix}} = 5$. The ratio is representative of ductal carcinoma in situ in glandular tissue in the breast under a strain of 5%, cf. [15]. The spatial distribution of the material is shown in Figure 3.4. The problem is $\Omega_0 = [0, L] \times [0, L]$ with $L = 10$ units. We employ a 10×10 FE mesh (for more fundamental detail about computational mechanics and finite element



(a) Problem configuration.



(b) Reference configuration of the material parameters E in log-scale.

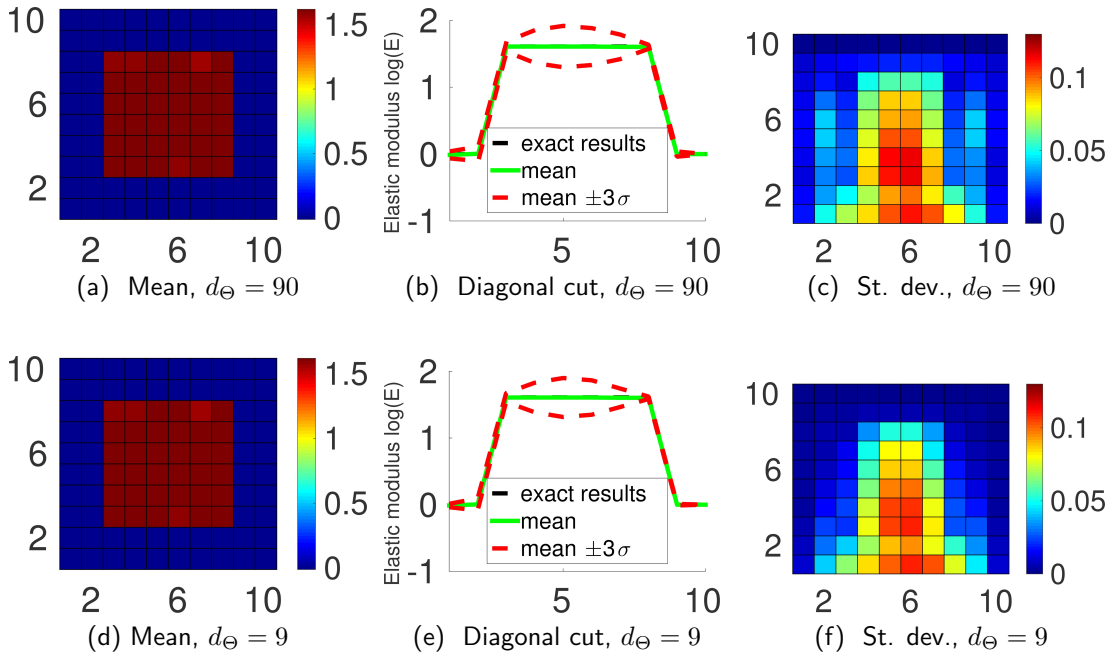
Figure 3.4: Problem and reference configuration.

methods we refer to Section 2.3 and the references therein). Displacement boundary conditions are employed which resemble those encountered when static pressure is applied on a tissue with the ultrasound transducer invoking a 1% strain as depicted in Figure 3.4. In particular, the boundary displacements at the bottom ($x_2 = 0$) are set to zero and at the top ($x_2 = 10$) the vertical displacements are set to -0.1 and the horizontal displacements equal to zero. The vertical edges ($x_1 = 0, 10$) are traction-free. The parameter values at the top row of elements are assumed known and equal to the exact values (E_{matrix}) otherwise any solutions for which $\frac{E_{inclusion}}{E_{matrix}} = 5$ would yield the same likelihood [43].[4] The interior observed displacements, generated using the reference configuration, were subsequently contaminated with Gaussian noise such that the resulting Signal-to-Noise Ratio (SNR) was $\mathrm{SNR} = 10^5$. We adopt a very vague prior, i.e., $\lambda_{0,1} = 10^{-10}$.

In the top row of Figure 3.5 various aspects of the posterior of the elastic moduli using 90 basis vectors, $d_\Theta = 90$ (equal to the total number of unknown material parameters, $d_\Psi = 90$), are depicted and are compared with the corresponding results $d_\Theta = 9$ (second row). One can see that the inferred posterior means are practically identical and coincide with the ground truth. The same can be said for the posterior variances which can be captured to a large extent by employing only $d_\Theta = 9$ reduced coordinates.

A more detailed comparison of the inferred posterior for various d_Θ is depicted in Figure 3.6. In the right subfigure also the relative information gain (as defined in

[4] This is only required for problems with Dirichlet boundary conditions as configurations with the same material parameter ratio results in the same displacements. Thus, the inversion scheme will identify the correct ratio of the material parameters but not necessarily the correct magnitude of it. This is not the case for given forces/pressure/stresses, Neumann boundary conditions.



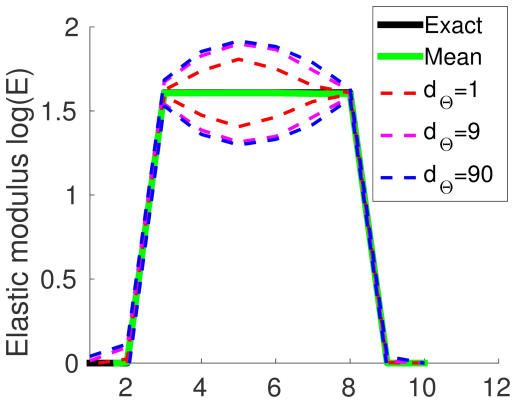
(a) Mean, $d_\Theta = 90$ (b) Diagonal cut, $d_\Theta = 90$ (c) St. dev., $d_\Theta = 90$

(d) Mean, $d_\Theta = 9$ (e) Diagonal cut, $d_\Theta = 9$ (f) St. dev., $d_\Theta = 9$

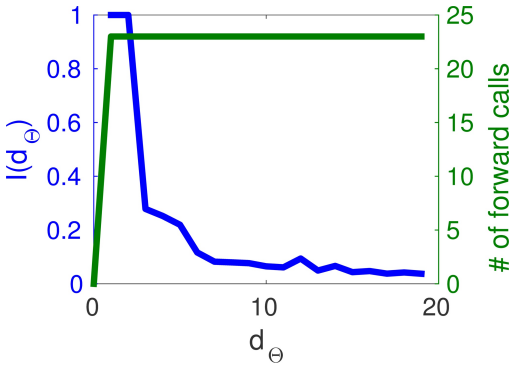
Figure 3.5: The first row corresponds to results derived with $d_\Theta = 90$ and the second row to $d_\Theta = 9$. Figures (a), (d) depict the posterior mean $\boldsymbol{\mu}$ of the elastic moduli E in log-scale which is shown to be independent of the number of reduced coordinates d_Θ. Figures (b), (e) show the posterior mean and posterior quantiles (± 3 standard deviations) along the diagonal from $(0, 0)$ to $(10, 10)$. Figures (c), (f) depict the posterior standard deviation. The differences are indistinguishable which implies that the full posterior ($d_\Theta = 90$) can be very well approximated with only $d_\Theta = 9$ reduced coordinates/basis vectors.

Section 3.2.5) and the number of forward calls (which determines the computational cost) as a function of the number of reduced coordinates/basis vectors is shown. One can notice that the information gain drops to relatively small values only after a small number of reduced coordinates (after the $d_\Theta = 6$, it drops below 10%). For the posterior approximation obtained with $d_\Theta = 9$ (which as shown earlier is practically indistinguishable from the full-order result with $d_\Theta = 90$) only 23 forward calls are needed. These forward calls are performed at $d_\Theta = 1$ and for additional reduced coordinates no further forward calls are required. A more detailed account of the optimization with regard to the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{W}$ can be seen in Figure 3.7, where the evolution of the corresponding variational objectives F_μ and F_W (Section 3.2.4) is plotted. We note again that the $\boldsymbol{\mu}$-updates are the only ones that require forward calls. The optimization results with regard to F_W are shown for $d_\Theta = 9$. These are performed using the Barzilai-Borwein step size selection discussed previously,

which results in a non-monotone but robust optimization.

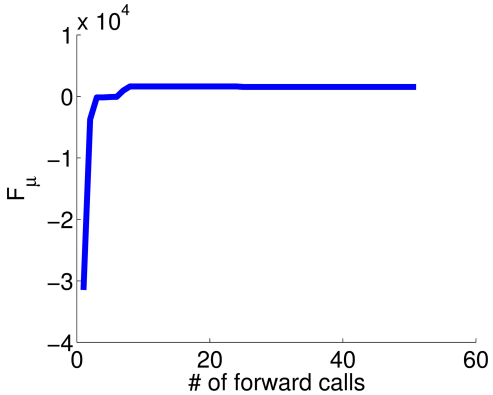


(a) Posterior mean and credible intervals at ± 3 standard deviations (dashed lines) along the diagonal from $(0, 0)$ to $(10, 10)$ for various values of d_Θ.

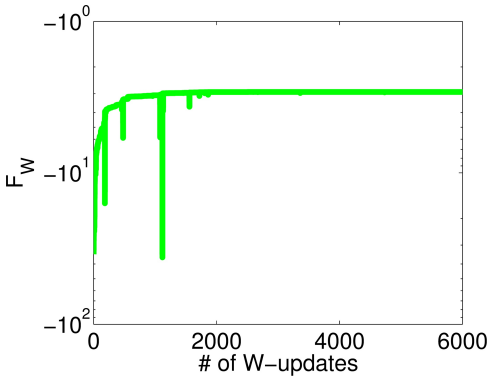


(b) Information gain $I(d_\Theta)$ — and computational cost — as measured by the number of forward calls over the number of dimensions, d_Θ.

Figure 3.6: Posterior statistics on the diagonal cut and information gain.



(a) F_μ over all bases



(b) F_W with 9 bases

Figure 3.7: (a): F_μ over the total number of $\boldsymbol{\mu}$-updates. (b): F_W during the $\boldsymbol{W}$-update, after adding the ninth basis.

The 9 most important basis vectors $\boldsymbol{w}_i$ can be seen in Figure 3.8, in decreasing order, based on the corresponding variance λ_i^{-1}.

Finally, the posterior of τ is depicted in Figure 3.9. One can observe that the magnitude is captured correctly, compared to the exact value, i.e., the corresponding variance of the Gaussian noise with which the data was contaminated.

The aforementioned results were verified by employing importance sampling, discussed in Section 3.2.6. The effective sample size (ESS, Equation (3.50)) was 0.25 (for

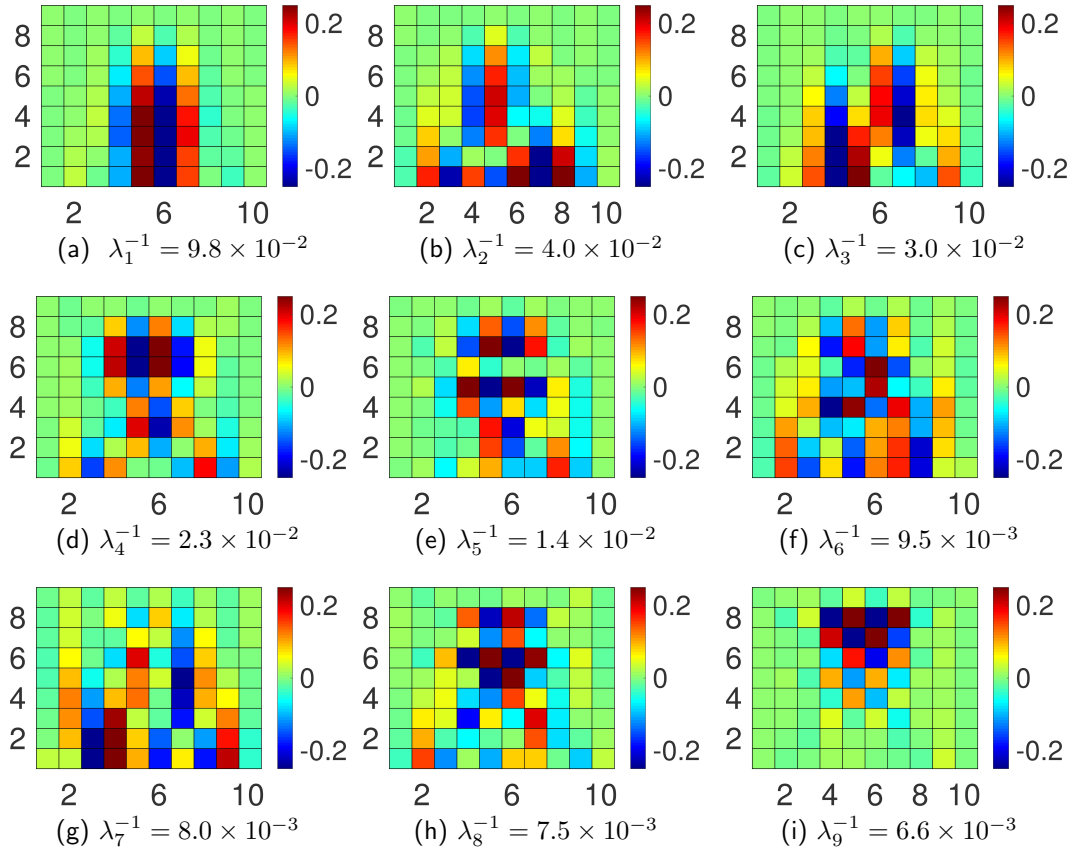


Figure 3.8: The first 9 basis vectors $\boldsymbol{w}_i$ in decreasing order, based on the corresponding variance λ_i^{-1}. One notes that the variance captured by the 9^{th} reduced coordinate is more than one magnitude smaller than that of the 1^{st} reduced coordinate.

$d_\Theta = 9$) which suggests that a good approximation to the actual posterior is provided by then VB result [128]. More importantly, in Figures 3.10 and 3.11 the first- and second-order statistics of the exact posterior (estimated with importance sampling) is shown and displays the good approximation of the posterior with VB.

Example 2: With an incompressible Mooney-Rivlin material

Nonlinear, hyperelastic models have been successfully used in the past to describe the behavior of several biomaterials [191, 192, 11]. In this example, we employ the Mooney-Rivlin model [193, 194] that is characterized by the following strain energy density function w (Equation (2.29)):

$$w = c_1(\hat{I}_1 - 3) + c_2(\hat{I}_2 - 3) + \frac{1}{2}\kappa(\log J)^2. \tag{3.54}$$

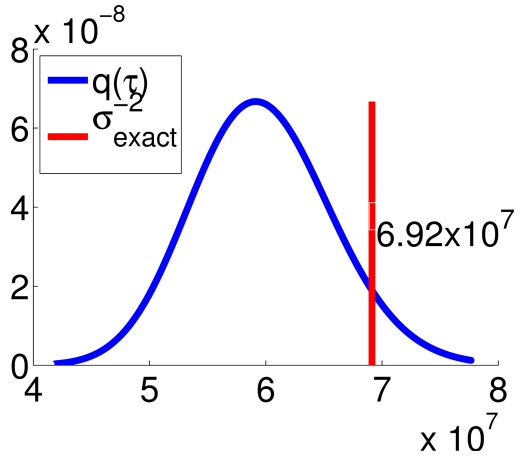
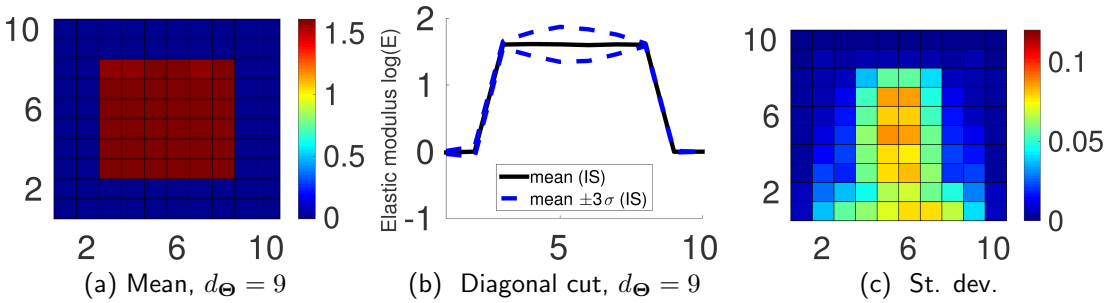


Figure 3.9: Posterior distribution $q(\tau)$ for 9 bases and the exact value.



(a) Mean, $d_{\boldsymbol{\Theta}} = 9$ (b) Diagonal cut, $d_{\boldsymbol{\Theta}} = 9$ (c) St. dev.

Figure 3.10: First- and second-order statistics of the exact posterior as estimated with importance sampling. Subfigure (a) depicts the posterior mean $\boldsymbol{\mu}$ of the elastic moduli E in log-scale. Subfigure (b) shows the posterior mean and posterior quantiles (± 3 standard deviations) along the diagonal from $(0, 0)$ to $(10, 10)$ and subfigure (c) pictures the posterior standard deviation. These should be compared with the VB approximations in Figure 3.5.

κ is the bulk modulus, $J = det(\boldsymbol{F})$ and $\hat{I}_1 = \frac{I_1}{J^{2/3}}, \hat{I}_2 = \frac{I_2}{J^{4/3}}$, where I_1, I_2 are the first and second invariants of the left Cauchy-Green deformation tensor $\boldsymbol{b} = \boldsymbol{F}\boldsymbol{F}^T$. The last term in Equation (3.54) is related to volumetric deformations whereas the first two terms to distortional. We consider here an *incompressible* material, i.e., $J = 1$, in which case the bulk modulus κ plays the role of a penalty parameter that enforces this constraint. We employ the three-field Hu-Washizu principle in order to enforce incompressibility and suppress volumetric locking [195, 144]. The three-field formulation requires a separate integration rule for the dilatational stiffness contribution. The bulk modulus is chosen as a function of c_1 with $\kappa = \kappa_o c_1$. We use $\kappa_0 = 1000$ [195, 196]. The higher κ_0 is, the stronger is the incompressibility constraint.

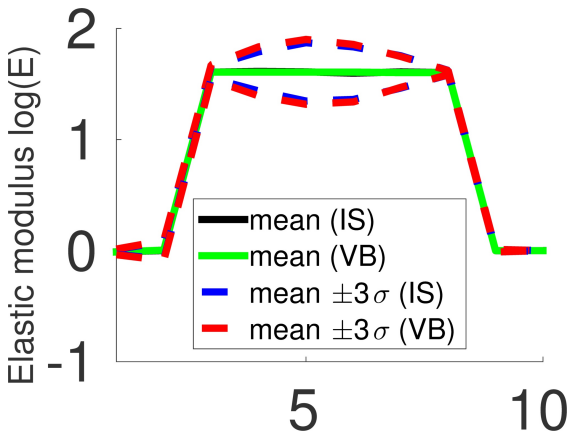


Figure 3.11: Posterior mean and posterior quantiles (± 3 standard deviations) along the diagonal from (0, 0) to (10, 10) for VB and importance sampling (IS).

In this example, we assume $c_2 = 0$, which reduces the model to an uncoupled version of the incompressible neo-Hookean model [141]. The remaining parameter c_1 is assumed to vary in the problem domain which can be seen in Figure 3.12. In this example we have two inclusions, an elliptic and a circular inclusion, with different material properties. In the larger, elliptic inclusion $c_1 = 4000$ (red), in the smaller, circular inclusion $c_1 = 3000$ (orange) and in the remaining material $c_1 = 1000$ (blue). The problem domain is $\Omega_0 = (0, L) \times (0, L)$ with $L = 50$. It is discretized with 200×200 finite elements of equal size and the governing equations are solved under plane strain conditions. The following boundary conditions are employed: both displacements are set to zero at the bottom ($x_2 = 0$) and vertical nodal loads $f = -100$ in the vertical direction (pointing downwards) is applied along the top, i.e., $x_2 = 50$. The vertical edges ($x_1 = 0, 50$) are traction-free.

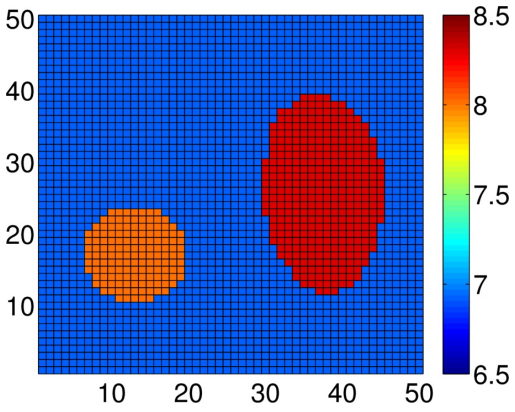


Figure 3.12: Reference c_1 distribution in the log-scale.

The forward model for the Bayesian identification employed a regular 50×50 mesh and only the corresponding (noisy) displacements at the nodes were used as data ($\hat{\boldsymbol{y}}$). We note that due to the different meshes employed the data will also contain model (discretization) errors. The SNRs reported in the sequence include also these errors. We further assumed that c_1 was constant within each of the elements which resulted in $d_\Psi = 2500$ *unknowns*. Using the displacements obtained from the fine 200×200 mesh we consider three settings:

- Case A (high SNR/low noise): without additional noise resulting in a SNR 1.93×10^3.
- Case B (medium SNR/medium noise): the data are contaminated with relatively smaller Gaussian noise resulting in a total SNR 1.89×10^3.
- Case C (small SNR/high noise): the data are contaminated with relatively larger Gaussian noise resulting in a total SNR 6.9×10^2.

The results presented in the sequence were obtained for $\lambda_{0,1} = 5 \times 10^{-1}$ and the material parameters are plotted in the log-scale.

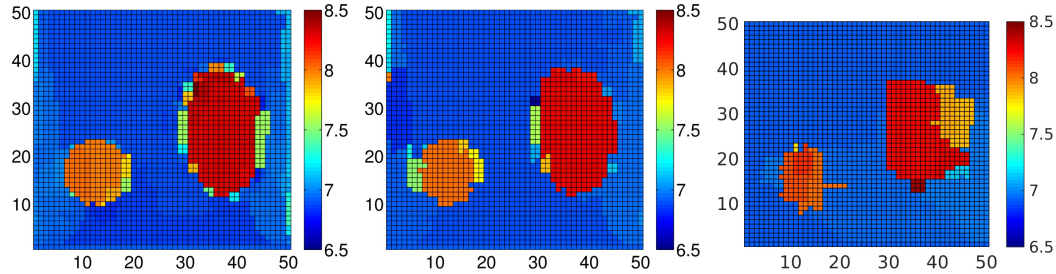


Figure 3.13: Posterior mean of c_1 in log-scale for Cases A (large SNR), B (medium SNR) and C (small SNR).

Figure 3.13 depicts the posterior mean $\boldsymbol{\mu}$ for the aforementioned three cases. Figure 3.14 displays the spatial distribution of the posterior standard deviation as obtained by using the reduced coordinates. We note that in all cases (low to high SNR), $\boldsymbol{\mu}$ provides a reasonable approximation of the ground truth. The advantage of the proposed as well as all Bayesian techniques is that probabilistic estimates can be obtained in the form of the posterior density. This is illustrated in Figure 3.15 which depicts the posterior along the diagonal from $(0, 0)$ to $(50, 50)$. Firstly, we note that in all cases the posterior quantiles envelop by-and-large the ground truth. Secondly, as expected, these credible intervals are larger in cases where the SNR is smaller (noise is larger). Thirdly and most importantly, we mark that these posterior approximations can be obtained by operating

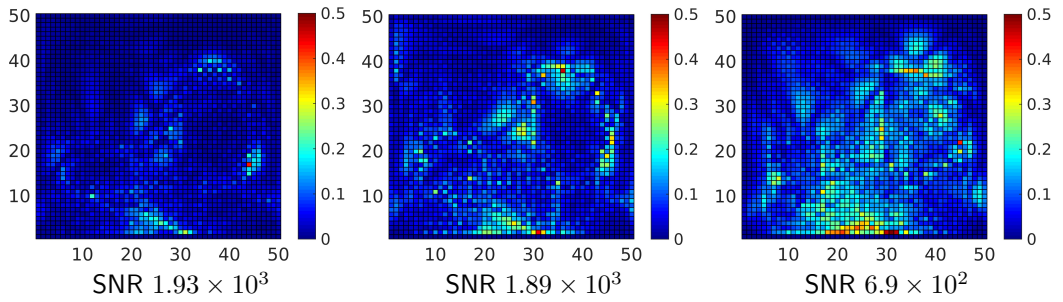


Figure 3.14: Posterior standard deviation of c_1 in log-scale for Cases A (large SNR, $d_\Theta = 10$), B (medium SNR, $d_\Theta = 12$) and C (small SNR, $d_\Theta = 13$).

on subspaces of dramatically reduced dimension in relation to the number of unknowns (2500).

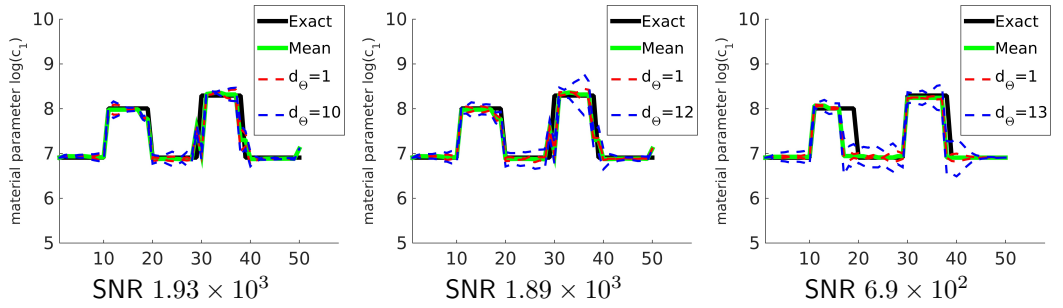


Figure 3.15: Posterior mean and credible intervals at ± 2 standard deviations (dashed lines) along the diagonal from $(0, 0)$ to $(50, 50)$ for various values of d_Θ and for Cases A (large SNR), B (medium SNR) and C (small SNR). The larger numbers of d_Θ correspond to the converged results as determined by Figure 3.16.

Figure 3.16 depicts the relative information gain $I(d_\Theta)$ (Section 3.2.5) for each SNR. The behavior of the information gain depends on the ratio of the prior $\lambda_{0,i}$ and the posterior of the variance λ_i. As with the previous example, it exhibits a relative quick decay after a small number of reduced coordinates have been added. Figure 3.16 shows also the number of forward calls as a function of d_Θ. As it was observed previously, the effort is expended in the beginning and in all cases the final result is obtained with less than 40 forward calls.

Figure 3.17 shows the evolution of F_μ as a function of the number of forward calls. Figure 3.18 depicts the corresponding evolution of F_W for $d_\Theta = 2$ and for all three SNR cases.

Finally, Figure 3.19 depicts 5 basis vectors $\boldsymbol{w}_i$ for each SNR in decreasing order

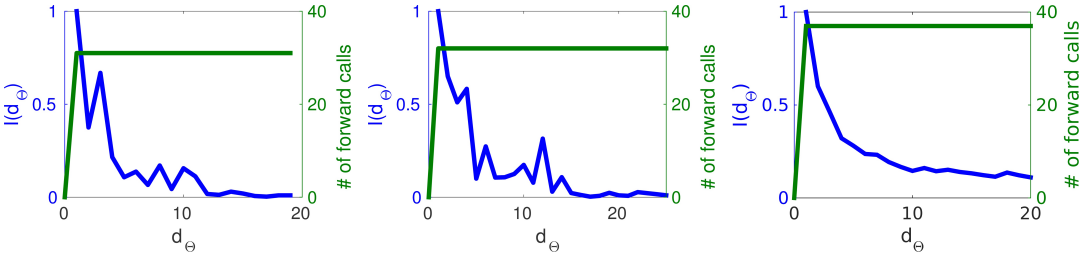


Figure 3.16: Information gain $I(d_\Theta)$ — and computational cost — as measured by the number of forward calls.

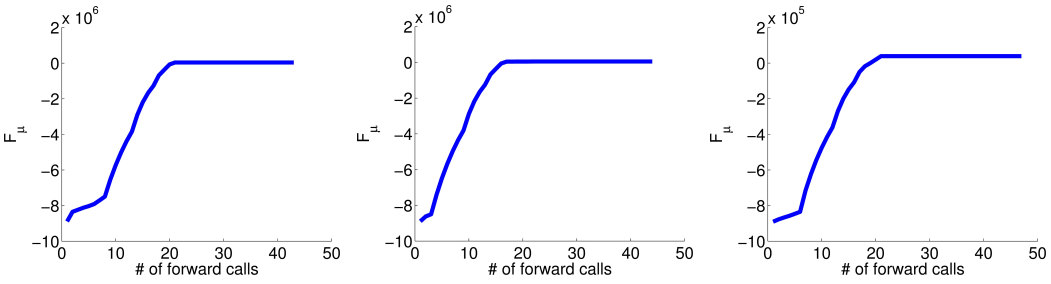


Figure 3.17: F_μ for the different SNR.

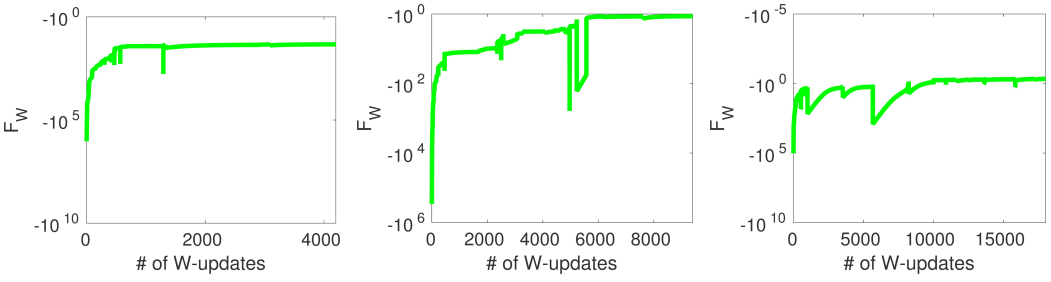


Figure 3.18: F_W for the different SNR and $d_\Theta = 2$.

based on the corresponding variance λ_i^{-1}. While similarities are observed, the basis vectors are not identical as compared across the three different noise levels, reflecting the fact that each dataset is informative along different directions in the $\mathbf{\Psi}$ space. However, it is clear that regions in the vicinity of (or within) the inclusions exhibit larger posterior variability. Also one expects, the associated variances are larger as the SNR is smaller (i.e., the noise level is higher).

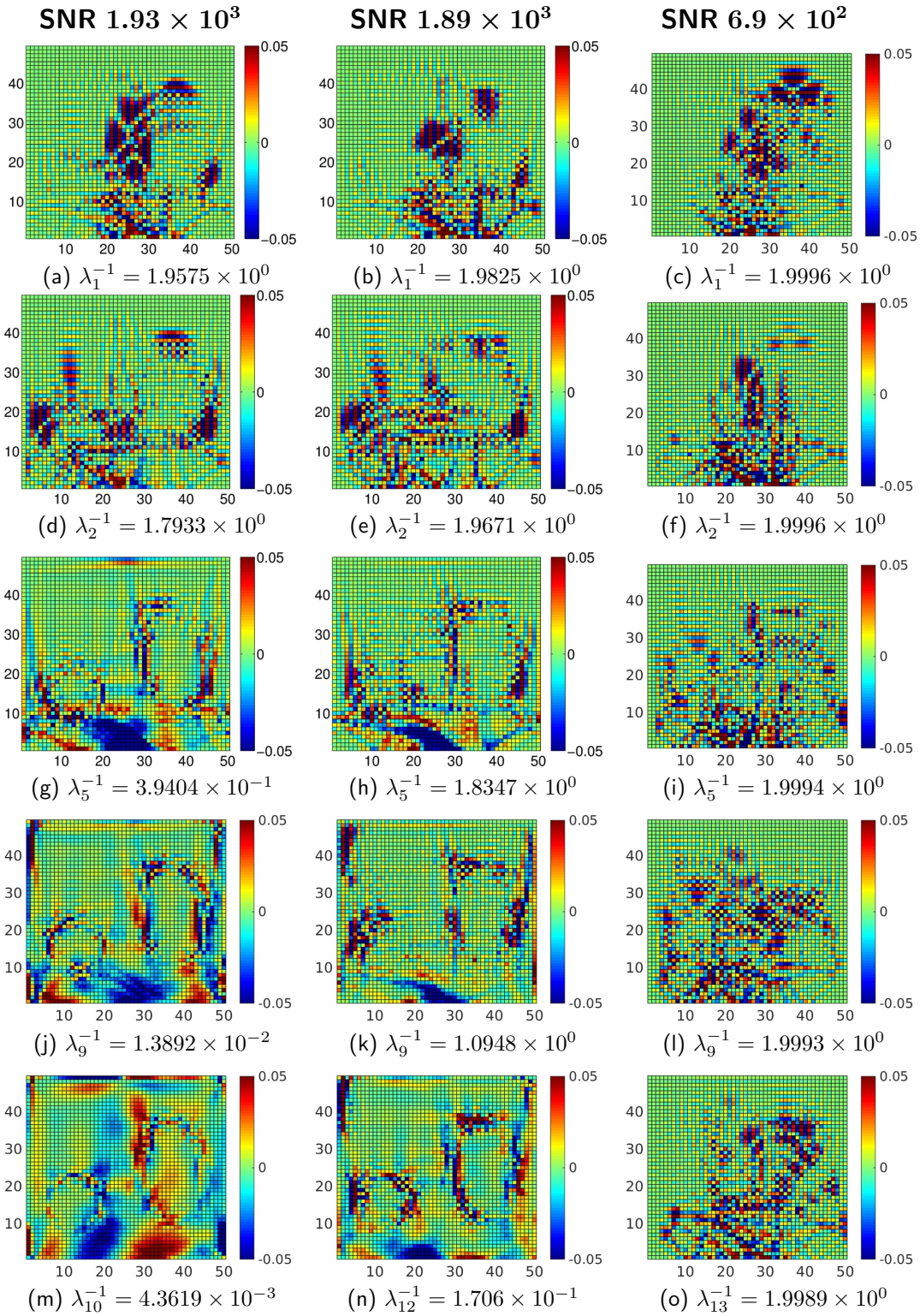
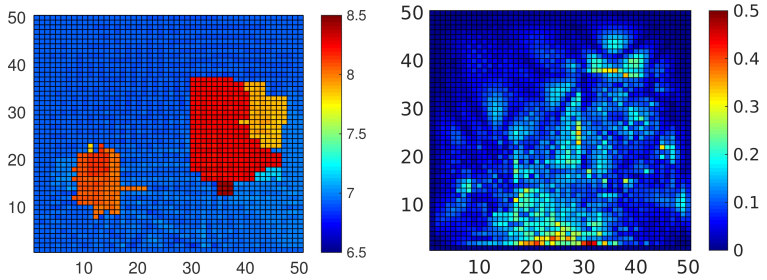


Figure 3.19: Some important selected basis vectors for Cases A (large SNR), B (medium SNR) and C (small SNR). The vectors are ordered based on decreasing variance λ_i^{-1}.

The aforementioned results for the largest noise case ($\mathrm{SNR} = 6.9 \times 10^2$) were verified by employing importance sampling as discussed in Section 3.2.6. The effective sample size (ESS, Equation (3.50)) was 0.15 (for $d_\Theta = 13$) which suggests a good approximation to the actual posterior is provided by the VB result [128]. More importantly, as it is shown in Figures 3.20 and 3.21, the first- and second-order statistics of the exact posterior (estimated with importance sampling) are very close to the ones computed with the VB approximation.



(a) Mean, $d_{\boldsymbol{\Theta}} = 13$ (b) St. dev., $d_{\boldsymbol{\Theta}} = 13$

Figure 3.20: First- and second-order statistics of the exact posterior ($\mathrm{SNR} = 6.9 \times 10^2$), as estimated with importance sampling. Subfigure (a) depicts the posterior mean of c_1 in log-scale. Subfigure (b) displays the posterior standard deviation. These should be compared with the VB approximations in Figure 3.13 and Figure 3.14.

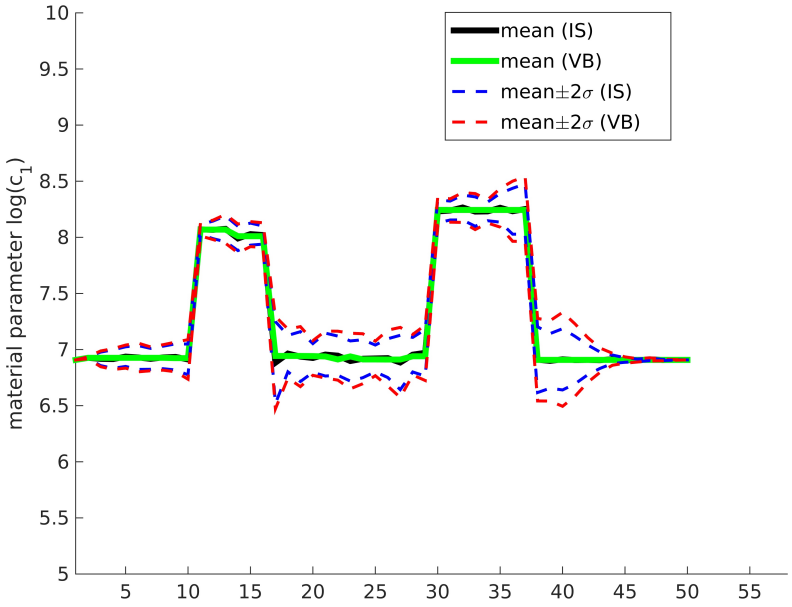


Figure 3.21: Posterior mean and posterior quantiles (± 2 standard deviations) along the diagonal from (0, 0) to (50, 50) for VB and importance sampling (IS).

3.4 Summary

We introduced a novel Variational Bayesian framework for the solution of nonlinear inverse problems and demonstrated its capabilities in problems in elastography. The main advantage of the proposed methodology is the ability to find a much lower-dimensional subspace where a good approximation to the exact posterior can be obtained. The identification of the reduced basis set is found on a fully Bayesian argumentation that employs Variational approximations to the exact posterior. Information-theoretic criteria have been proposed in order to adaptively identify the cardinality of the reduced coordinates. The posterior approximations are obtained with a limited number of calls to the forward solver. For the computation of the response and its derivatives (in all problems considered fewer than 40 such calls were needed). Furthermore, with the use of importance sampling, the (minute) bias in the posterior estimates can be corrected and consistent statistics of the *exact posterior* can be estimated.

A possibility that could further reduce the computational effort is the use of forward solvers operating on a hierarchy of resolutions. However, in these approaches location-dependent model error needs to be considered. This will further complicate the Bayesian variational approach. Starting with the coarsest (and less expensive) model some of the features of the posterior can be obtained with minimal cost and these can be further refined by a smaller number of calls to finer resolution solvers. The resolution of the forward model could also be adaptively altered in regions where the posterior variance appears to be larger. Obtaining efficiently, accurate and fully-Bayesian solutions is a critical step in enabling the use of model-based techniques on a patient-specific basis for medical diagnosis.

Another possible extension is the usage of *mixtures* of Gaussian densities in order to provide better approximations to highly non-Gaussian or even *multimodal* posteriors [93]. Such situations arise frequently in cases where very sparse and/or very noisy data is available and represent the most challenging setting for associated inverse problems [59]. Tools along the aforementioned lines, offer appealing possibilities for identifying multiple low-dimensional subspaces and associated basis vectors which *locally* provide good posterior approximations and when combined, offer an accurate global solution.

Chapter 4

Multimodal, high-dimensional, model-based, Bayesian inverse problems with applications in biomechanics

> *Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable.*
>
> Bobby Bragan, 1917-2010.

This chapter is based on the publication: I. M. Franck, P.S. Koutsourelakis, *Multimodal, high-dimensional, model-based, Bayesian inverse problems with applications in biomechanics*, Journal of Computational Physics, Volume 329, 15 January 2017, Pages 91-125 [197]

4.1 Why it can be that important to capture multimodality

Many posteriors are multimodal and highly non-Gaussian, which is another challenge we want to address in this thesis. That is, the identification of multiple posterior modes. In the context of elastography, multimodality can originate from anisotropic materials [198], wrong/missing information from images/measurements [199] or the imaging modality employed [200]. In all cases, each posterior mode can lead to different diagnostic conclusions. Therefore, it is very important to identify them and correctly assess their posterior probabilities. The majority of Bayesian strategies for the solution of computationally intensive inverse problems operate under the assumption of a unimodal posterior or focuses on the approximation of a single mode of the posterior. Some numerical inference tools, based on SMC or other tempering mechanisms [87, 88, 89], have been developed but require a very large number of forward model calls particularly when the dimension of unknowns increases. We finally note that the treatment of multimodal densities in high-dimensions has attracted significant interest in atomistic simulations in the context of free energy computations [201, 202], but in such problems (apart from other distinguishing features) the cost per density evaluation (i.e., one MD time-step) is smaller than in our setting.

In this chapter, we discuss a Variational Bayesian (VB) strategy that extends our work of the previous chapter. Therein, we have shown how accurate approximations of the true posterior can be attained by identifying a low-dimensional subspace where posterior uncertainty is concentrated. This has led to computational schemes that require only a few tens of forward model runs in the problems investigated. Nevertheless, the previous chapter was based on the assumption of a unimodal posterior which we propose overcoming in this chapter by employing a mixture of multivariate Gaussians. Mixture models have also been employed in various statistics and machine learning applications (e.g., speaker identification [91], data clustering [92]) and also in combination with Variational Bayesian inference techniques [93, 79, 94]. We note that a different VB strategy that also makes use of mixtures of Gaussians to solve model-based inverse problems has been proposed in [109]. Nonetheless, all the presented problems have inexpensive likelihoods, relatively low-dimensions and multiple data/measurements. In contrast, the inverse problems considered here are based on a single experiment, a single observation vector and a large number of unknown latent variables.

Within the novel Bayesian framework, which integrates a dimensionality reduction for each mixture component, we propose an adaptive algorithm based on information-theoretic criteria for the identification of the number of the required mixture components (Section 4.2). Furthermore, we present the parametrization of the proposed model in Section 4.2 where we also discuss a Variational-Bayesian Expectation-Maximization [79] scheme for performing inference and learning. In Section 4.3, we

present numerical illustrations involving a simple toy-example and an example in the context of elastography.

4.2 Methods

This section discusses the advocated methodological framework. In Section 4.2.1, we present a Bayesian mixture model that can identify lower-dimensional subspaces where most of the posterior mass is concentrated as well as accounting for multiple modes. The prior assumptions for the model parameters are summarized in Section 4.2.2. In Section 4.2.3, we discuss a Variational Bayesian Expectation-Maximization scheme for computing efficiently approximations to the posterior for a fixed number of mixture components. In Section 4.2.4, we examine a scheme for determining the appropriate number of such components. Finally, in Section 4.2.5, we discuss how to assess the accuracy of the computed approximation as well as a way to correct for any bias if this is deemed to be necessary.

Canonical formulations of model-based, inverse problems postulate the existence of a forward model that typically arises from the discretization of governing equations, such as PDEs/ODEs. The following discussions are based on the forward model discussed in Section 3.1 with the likelihood from Equation (3.2)

$$p(\hat{\boldsymbol{y}}|\boldsymbol{\Psi},\tau) = \left(\frac{\tau}{2\pi}\right)^{d_y/2} e^{-\frac{\tau}{2}||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\Psi})||^2}. \tag{4.1}$$

The intractability of the map $\boldsymbol{y}(\boldsymbol{\Psi})$ precludes the availability of closed-form solutions for the posterior and necessitates the use of various sampling schemes, such as those discussed in the introduction. This task is seriously impeded by a) the need for repeated solutions of the discretized forward problem of which each can be quite computationally taxing, b) the high dimensionality of the vector of unknowns $\boldsymbol{\Psi}$ which hinders the efficient search (e.g., by sampling) of the latent parameter space and further increases the computational burden. Within the previous chapter, we alleviate these difficulties by proposing adequate approximations and dimensionality-reduction techniques that are seamlessly integrated in the inference framework. Within this section, we attempt to overcome well-known limitations that have to do with the multimodality of the posterior and which further exacerbate these problems. Multimodality is inherently related to the ill-posedness of inverse problems and its potential can increase when the dimension of the vector of unknowns increases and/or the noise is amplified.

4.2.1 Bayesian mixture model

In this section and in view of the aforementioned desiderata, we propose the augmented formulation of the Bayesian inverse problem.

- In order to capture multiple modes of the posterior (if those are present) we introduce the *discrete, latent variable* s which takes integer values between 1 and S. The latter represents the number of modes identified, each of which will be modeled with a multivariate Gaussian. The cardinality of the model, i.e., S is learned in a manner that is described in the sequel.

- In order to identify a lower-dimensional representation of the unknowns $\boldsymbol{\Psi}$ we define the latent variables $\boldsymbol{\Theta} \in \mathbb{R}^{d_\Theta}$ such that $d_\Theta \ll d_\Psi$. The premise here is that while $\boldsymbol{\Psi}$ is high-dimensional, its posterior can be adequately represented on a subspace of dimension d_Θ that captures most of the variance. As we have argued in Chapter 3 these latent variables can give rise to a PCA-like decomposition of the form:

$$\boldsymbol{\Psi} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\Theta} + \boldsymbol{\eta}, \tag{4.2}$$

  where $\boldsymbol{\mu} \in \mathbb{R}^{d_\Psi}$ is the mean vector and the columns of the orthogonal matrix $\boldsymbol{W} \in \mathbb{R}^{d_\Psi \times d_\Theta}$ ($\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_{d_\Theta}$) span the aforementioned subspace with reduced coordinates $\boldsymbol{\Theta}$. The vector $\boldsymbol{\eta} \in \mathbb{R}^{d_\Psi}$ captures the residual variance (noise) that complements the main effects.

  In view of multimodal posteriors and since each mode implies a different mean and a different lower-dimensional subspace (Figure 4.1), we advocate in this chapter S expansions of the form:

$$\boldsymbol{\Psi}_s = \boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}, \quad s = 1, 2, \ldots, S. \tag{4.3}$$

  The notation $\boldsymbol{\Psi}_s$ implies the representation of $\boldsymbol{\Psi}$ within mode s. The orthogonal matrices $\boldsymbol{W}_s$ denote the potentially different subspaces associated with each of the modes. In principle, the dimension d_Θ of the reduced subspaces can also vary with s, but we do not consider this here for simplicity of notation.

We distinguish in the following between *latent variables*: s, $\boldsymbol{\Theta}$, $\boldsymbol{\eta}$ and τ, and model parameters: $\boldsymbol{\mu} = \{\boldsymbol{\mu}_j\}_{j=1}^S$, $\boldsymbol{W} = \{\boldsymbol{W}_j\}_{j=1}^S$. We seek point estimates for the latter due to their high dimension (of the order of $d_\Psi \gg 1$) and (approximations) of the actual (conditional) posterior for the former. Based on the parametrization adopted, the likelihood of Equation (3.2) takes the form:

$$p(\hat{\boldsymbol{y}}|s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{\mu}_s, \boldsymbol{W}_s) \propto \tau^{d_y/2} e^{-\frac{\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta})||^2}. \tag{4.4}$$

A graphical illustration of the proposed probabilistic generative model is shown in Figure 4.2.

Following the standard Bayesian formalism, one would complement the aforementioned likelihood with priors on the model parameters $p(\boldsymbol{\mu}, \boldsymbol{W})$ and the latent variables $p(\boldsymbol{\Theta}, \boldsymbol{\eta}, s, \tau)$, in order to obtain the joint posterior (given S):

$$p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{\mu}, \boldsymbol{W}|\hat{\boldsymbol{y}}) \propto \; p(\hat{\boldsymbol{y}}|s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{\mu}_s, \boldsymbol{W}_s)\; p(\boldsymbol{\Theta}, \boldsymbol{\eta}, s, \tau)\; p(\boldsymbol{\mu}, \boldsymbol{W}). \tag{4.5}$$

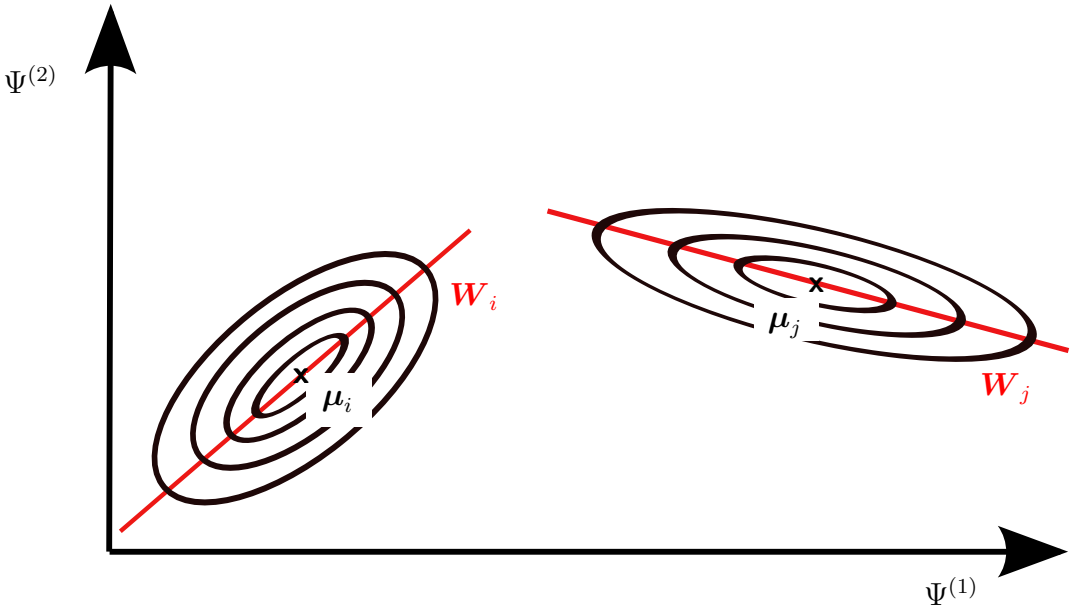


Figure 4.1: Illustration of the multimodal representation for $S = 2$ in 2D, i.e., when $\boldsymbol{\Psi} = \{\Psi^{(1)}, \Psi^{(2)}\}$.

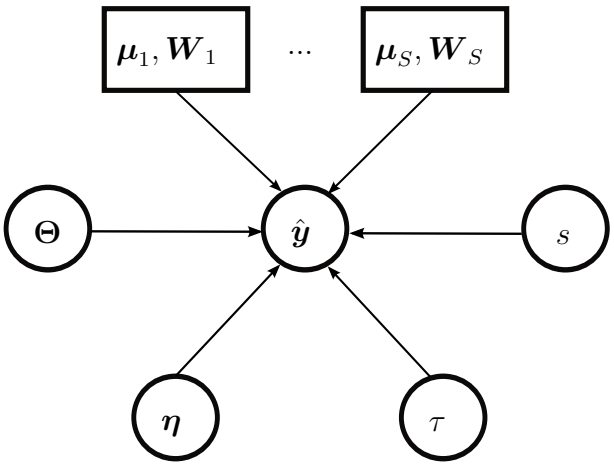


Figure 4.2: Graphical representation of the proposed generative probabilistic model. Circles denote random variables, solid rectangles, model parameters and arrows denote dependencies [124].

We discuss the specific form of the priors (and associated hyperparameters) in Subsection 4.2.2 as well as the inference/learning strategy we propose in 4.2.3. However, we

note at this stage that given this posterior, one would obtain a *mixture* representation of the unknown material parameters $\boldsymbol{\Psi}$, as implied by Equation (4.3). In particular, given values for $(\boldsymbol{\mu}, \boldsymbol{W}) = \{\boldsymbol{\mu}_j, \boldsymbol{W}_j\}_{j=1}^{S}$, it directly follows that the *posterior* $p(\boldsymbol{\Psi}|\boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}})$ (given S) of $\boldsymbol{\Psi}$ is:

$$\begin{aligned} & p(\boldsymbol{\Psi}|\, \boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}}) \\ & \quad = \textstyle\sum_{s=1}^{S} \int p(\boldsymbol{\Psi}, s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, |\boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}})\; d\boldsymbol{\Theta}\; d\boldsymbol{\eta}\; d\tau \\ & \quad = \textstyle\sum_{s=1}^{S} \int p(\boldsymbol{\Psi}|s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{\mu}_s, \boldsymbol{W}_s)\; p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, |\boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}})\; d\boldsymbol{\Theta}\; d\boldsymbol{\eta}\; d\tau \\ & \quad = \textstyle\sum_{s=1}^{S} \int \delta(\boldsymbol{\Psi} - (\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}))\; p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, |\boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}})\; d\boldsymbol{\Theta}\; d\boldsymbol{\eta}\; d\tau, \end{aligned} \tag{4.6}$$

where the conditional posterior $p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, |\boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}})$ is found from Equation (4.5). We discuss in Section 4.2.3 how the posterior on the latent variables is approximated as well as the values (point estimates) for the model parameters $\boldsymbol{\mu}, \boldsymbol{W}$ are computed.

4.2.2 Prior specification for mixture model with dimensionality reduction

We assume that, a priori, the precision τ of the observation noise is independent of the remaining latent variables $\boldsymbol{\Theta}, \boldsymbol{\eta}, s$, i.e.:

$$p(\boldsymbol{\Theta}, \boldsymbol{\eta}, s, \tau) = p(\boldsymbol{\Theta}, \boldsymbol{\eta}, s)\; p_\tau(\tau). \tag{4.7}$$

In particular, we employ:

- a Gamma prior on τ: as in the previous chapter, cf. Equation (3.3), we employ a (conditionally) conjugate Gamma prior $p_\tau(\tau)$:

  $$p_\tau(\tau) \equiv Gamma(a_0, b_0). \tag{4.8}$$

  We use $a_0 = b_0 = 0$ which results in a non-informative Jeffreys' prior that is scale-invariant.

- We assume that $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ are a priori, *conditionally independent*, i.e., $p(\boldsymbol{\Theta}, \boldsymbol{\eta}, s) = p(\boldsymbol{\Theta}, \boldsymbol{\eta}|s)\; p_s(s) = p_\Theta(\boldsymbol{\Theta}|s)\; p_\eta(\boldsymbol{\eta}|s)\; p_s(s)$. We discuss each of these terms below:

  - We assume that each component s is, a priori, equally likely, which implies:

    $$p_s(s) = \frac{1}{S}, \quad s \in [1:S]. \tag{4.9}$$

    Hierarchical priors can readily be adopted (e.g., [79]), but we consider here the simplest possible scenario. An interesting extension would involve infinite models with Dirichlet Process priors [203, 204] which would enable the number of components S to be automatically determined. In this work, a less elegant but quite effective adaptive scheme for determining S is proposed in Section 4.2.4.

– A Gaussian prior on $\boldsymbol{\Theta}$:
The role of the latent variables $\boldsymbol{\Theta}$ is to capture the most significant variations of $\boldsymbol{\Psi}_s$ around its mean $\boldsymbol{\mu}_s$ as in Section 3.2.3 and 3.2.4. By significant we mean the directions along which the largest posterior uncertainty is observed. These represent the reduced coordinates along the subspace spanned by the column vectors of $\boldsymbol{W}_j$. We assume therefore, cf. Equation (3.19), that a priori, these are independent, have zero mean and follow a multivariate Gaussian:

$$p_\Theta(\boldsymbol{\Theta}|s) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_{0,s}^{-1}), \tag{4.10}$$

where $\boldsymbol{\Lambda}_{0,s} = diag(\lambda_{0,s,i})$, $i = 1, ..., d_\Theta$ express prior variances along each of the latent principal directions.

– A Gaussian prior on $\boldsymbol{\eta}$:
As the role of these latent variables is to capture any residual variance (that is not accounted for by $\boldsymbol{\Theta}$), we assume that, *a priori*, $\boldsymbol{\eta}$ can be modeled by a multivariate Gaussian that has zero mean and an isotropic covariance (see Equation (3.20)):

$$p_\eta(\boldsymbol{\eta}|s) = \mathcal{N}(\mathbf{0}, \lambda_{0,\eta,s}^{-1}\boldsymbol{I}_{d_\Psi}). \tag{4.11}$$

For the model parameters $\boldsymbol{\mu}, \boldsymbol{W}$, we assume that, a priori, the parameters associated with each component $j = 1, \ldots, S$ are independent. In particular:

- Prior on each $\boldsymbol{\mu}_j$ for $j \in 1 : S$:
In general such priors must encapsulate not only the information/beliefs available a priori to the analyst but also reflect the physical meaning of the parameters $\boldsymbol{\Psi}$. We are motivated by applications in elastography where the goal is to identify inclusions that correspond to tumors and generally have very different properties from the surrounding tissue [15, 16]. The vector $\boldsymbol{\Psi}$ represents the spatial discretization of the material parameters, i.e., each of its entries corresponds to the value of the material parameter at a certain point in the physical domain. This structure is inherited by $\boldsymbol{\mu}_j$ and for this reason we employ a hierarchical prior that penalizes jumps between neighboring locations (on the spatial domain, [205]) in a manner controlled by appropriately selected hyperparameters. The model was discussed in detail in Section 3.2.3 and is now extended to each mixture component, the prior and hyperprior for each mixture component j follows with:

$$p(\boldsymbol{\mu}_j|\boldsymbol{\Xi}_j) \propto |\boldsymbol{\Xi}_j|^{1/2} e^{-\frac{1}{2}\boldsymbol{\mu}_j^T \boldsymbol{L}^T \boldsymbol{\Xi}_j \boldsymbol{L} \boldsymbol{\mu}_j}, \tag{4.12}$$

$$p(\boldsymbol{\Xi}_j) = \prod_{m=1}^{d_L} Gamma(a_\xi, b_\xi). \tag{4.13}$$

- Prior specification on each $\boldsymbol{W}_j$ for $j \in 1:S$:
  We require that each $\boldsymbol{W}_j$ is orthonormal, i.e., $\boldsymbol{W}_j^T \boldsymbol{W}_j = \boldsymbol{I}_{d_\Theta}$, where $\boldsymbol{I}_{d_\Theta}$ is the d_Θ-dimensional identity matrix. This is equivalent to employing a uniform prior on the Stiefel manifold $V_{d_\Theta}(\mathbb{R}^{d_\Psi})$, as discussed in Section 3.2.3.

4.2.3 Variational approximation

We note that inference (exact or approximate) for all the model parameters described previously would pose a formidable task particularly with regard to $\boldsymbol{\mu}$ and $\boldsymbol{W}$ which are of dimension of order $d_\Psi \gg 1$ in their larger dimension. For that purpose, we advocate a hybrid approach whereby Maximum-A-Posteriori (MAP) point estimates of the high-dimensional parameters $\boldsymbol{T} = (\boldsymbol{\mu}, \boldsymbol{W}) = \{\boldsymbol{\mu}_j, \boldsymbol{W}_j\}_{j=1}^S$ are obtained and the posterior of the remaining (latent) variables $s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau$ is approximated. To that end, we make use of the Variational Bayesian Expectation-Maximization scheme (VB-EM, [79] and Section 3.2.2) which provides a lower bound $\mathcal{F}$ on the log of the marginal posterior of $\boldsymbol{T} = (\boldsymbol{\mu}, \boldsymbol{W})$. This can be iteratively maximized by a generalized coordinate ascent (Figure 4.3) which alternates between finding optimal approximations $q(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ of the exact (conditional) posterior $p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau | \hat{\boldsymbol{y}}, \boldsymbol{T})$ and optimizing with respect to $\boldsymbol{T}$.

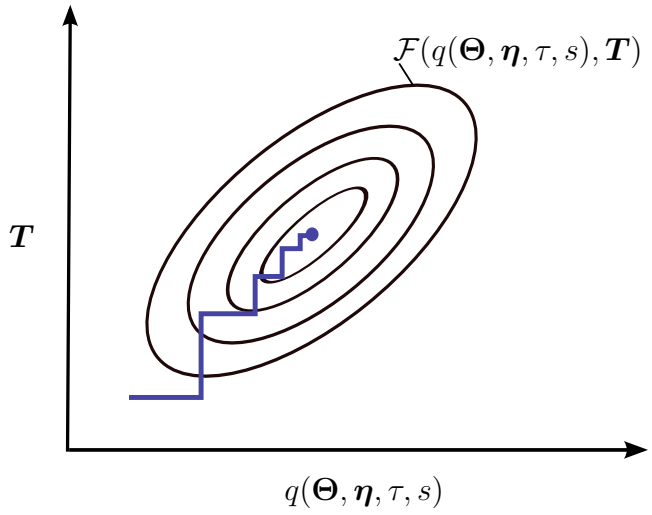


Figure 4.3: Schematic illustration of the advocated Variational Bayesian Expectation-Maximization for mixture of Gaussians (VB-EM, [79]).

On the basis of the discussion above and the separation between latent variables $(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ and model parameters $\boldsymbol{T}$, we can rewrite Equation (4.5) (for a given S) as follows:

$$p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{T} | \hat{\boldsymbol{y}}) = \frac{p(\hat{\boldsymbol{y}} | s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{T})\ p_s(s)\ p_\Theta(\boldsymbol{\Theta}|s)\ p_\eta(\boldsymbol{\eta}|s)\ p_\tau(\tau)\ p_T(\boldsymbol{T})}{p(\hat{\boldsymbol{y}})}. \quad (4.14)$$

We note that both sides of the equation above depend implicitly on S, i.e., the total number of components in the model. This is especially important for the model evidence term $p(\hat{\boldsymbol{y}})$ which we discuss in Section 4.2.4. We nevertheless omit S from the expressions in order to simplify the notation.

Furthermore, the *conditional* posterior of $(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau)$ given $\boldsymbol{T}$ is:

$$p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau | \boldsymbol{T}, \hat{\boldsymbol{y}}) = \frac{p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, \boldsymbol{T} | \hat{\boldsymbol{y}})}{p(\boldsymbol{T} | \hat{\boldsymbol{y}})}, \tag{4.15}$$

where $p(\boldsymbol{T}|\hat{\boldsymbol{y}})$ is the (marginal) posterior of the model parameters $\boldsymbol{T}$.

For an arbitrary density $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$ and by employing Jensen's inequality, it can be shown that (cf. Equation (3.8)):

$$\begin{aligned} \log p(\boldsymbol{T}|\hat{\boldsymbol{y}}) &= \log \textstyle\sum_{s=1}^{S} \int p(\boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}}) \; d\boldsymbol{\Theta} \; d\boldsymbol{\eta} \; d\tau \\ &= \log \textstyle\sum_{s=1}^{S} \int q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s) \frac{p(\boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)} \; d\boldsymbol{\Theta} \; d\boldsymbol{\eta} \; d\tau \\ &\geq \textstyle\sum_{s=1}^{S} \int q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s) \log \frac{p(\boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)} \; d\boldsymbol{\Theta} \; d\boldsymbol{\eta} \; d\tau \\ &= \mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s), \boldsymbol{T}). \end{aligned} \tag{4.16}$$

We note here that the variational lower bound $\mathcal{F}$ has a direct connection with the Kullback-Leibler (KL) divergence between $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$ and the (conditional) posterior $p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \boldsymbol{T}, \hat{\boldsymbol{y}})$. In particular, if we denote by $< \, . \, >_q$ expectations with respect to q, then:

$$\begin{aligned} KL\left(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s) || p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}}, \boldsymbol{T})\right) &= -\left\langle \log \tfrac{p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}}, \boldsymbol{T})}{q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)} \right\rangle_q \\ &= -\left\langle \log \tfrac{p(\boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}})}{p(\boldsymbol{T} | \hat{\boldsymbol{y}}) \; q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)} \right\rangle_q \\ &= \log p(\boldsymbol{T} | \hat{\boldsymbol{y}}) - \mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s), \boldsymbol{T}). \end{aligned} \tag{4.17}$$

The Kullback-Leibler divergence is by definition non-negative and becomes zero when $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s) \equiv p(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s | \hat{\boldsymbol{y}}, \boldsymbol{T})$. Hence, for a given $\boldsymbol{T}$, constructing a good approximation to the conditional posterior (in the KL divergence sense) is equivalent to maximizing the lower bound $\mathcal{F}(q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s), \boldsymbol{T})$ with respect to $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$. Analogously, maximizing $\mathcal{F}$ with respect to $\boldsymbol{T}$ (for a given $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$ leads to (sub-)optimal MAP estimates. This suggests an iterative scheme that alternates between:

- **VB-Expectation** step: Given the current estimate of $\boldsymbol{T}$, find the $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$ that maximizes $\mathcal{F}$.
- **VB-Maximization** step: Given the current $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$, find $\boldsymbol{T}$ that maximizes $\mathcal{F}$.

As in standard EM schemes [135], relaxed versions of the aforementioned partial optimization problems can be considered that improve upon the current $\mathcal{F}$ rather than finding the optimum at each iteration.

Using Equation (4.14), the lower bound $\mathcal{F}$ can be expressed as:

$$\begin{aligned}\mathcal{F}(q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s),\boldsymbol{T}) &= \left\langle \log \frac{p(\hat{\boldsymbol{y}}|s,\boldsymbol{\Theta},\boldsymbol{\eta},\tau,\boldsymbol{T})\ p_s(s)\ p_\Theta(\boldsymbol{\Theta}|s)\ p_\eta(\boldsymbol{\eta}|s)\ p_\tau(\tau)\ p_T(\boldsymbol{T})}{p(\hat{\boldsymbol{y}})\ q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s)} \right\rangle_q \\ &= \left\langle \log \frac{p(\hat{\boldsymbol{y}}|s,\boldsymbol{\Theta},\boldsymbol{\eta},\tau,\boldsymbol{T})\ p_s(s)\ p_\Theta(\boldsymbol{\Theta}|s)\ p_\eta(\boldsymbol{\eta}|s)\ p_\tau(\tau)}{q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s)} \right\rangle_q \\ &+ \log p_T(\boldsymbol{T}) - \log p(\hat{\boldsymbol{y}}) \\ &= \hat{\mathcal{F}}(q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s),\boldsymbol{T}) + \log p_T(\boldsymbol{T}) - \log p(\hat{\boldsymbol{y}}).\end{aligned} \tag{4.18}$$

We will omit the term $-\log p(\hat{\boldsymbol{y}})$ as it does not depend on q nor $\boldsymbol{T}$. It is apparent that the challenging term in $\hat{\mathcal{F}}$ involves the likelihood, i.e.:

$$\begin{aligned}\hat{\mathcal{F}}(q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s),\boldsymbol{T}) &= \left\langle \log \frac{p(\hat{\boldsymbol{y}}|s,\boldsymbol{\Theta},\boldsymbol{\eta},\tau,\boldsymbol{T})\ p_s(s)\ p_\Theta(\boldsymbol{\Theta}|s)\ p_\eta(\boldsymbol{\eta}|s)\ p_\tau(\tau)}{q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s)} \right\rangle_q \\ &= \left\langle \frac{d_y}{2}\log\tau - \frac{\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta})||^2 \right\rangle_q \\ &+ \left\langle \log \frac{p_s(s)\ p_\Theta(\boldsymbol{\Theta}|s)\ p_\eta(\boldsymbol{\eta}|s)\ p_\tau(\tau)}{q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s)} \right\rangle_q .\end{aligned} \tag{4.19}$$

The intractability of the map $\boldsymbol{y}(.)$ precludes an analytic computation of the expectation with respect to q, let alone the optimization with respect to this. While stochastic approximation techniques in the context of VB inference have been suggested [206] to carry out this task, these would require repeated forward solves (i.e., evaluations of $\boldsymbol{y}(.)$) which would render them impractical. For that purpose, as in Section 3.2.2, we invoke an *approximation* by using a first-order Taylor series expansion of $\boldsymbol{y}$ (given s) at $\boldsymbol{\mu}_s$, i.e.:

$$\boldsymbol{y}(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}) = \boldsymbol{y}(\boldsymbol{\mu}_s) + \boldsymbol{G}_s\left(\boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}\right) + \mathcal{O}(||\boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}||^2), \tag{4.20}$$

where $\boldsymbol{G}_s = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}|_{\boldsymbol{\Psi}=\boldsymbol{\mu}_s}$ is the gradient of the map at $\boldsymbol{\mu}_s$. We will discuss rigorous validation strategies of the approximation error thus introduced in Section 2.2.3. Truncating Equation (4.20) to first-order, the term $||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta})||^2$ in the exponent of the likelihood becomes:

$$\begin{aligned}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta})||^2 &= ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s) - \boldsymbol{G}_s\boldsymbol{W}_s\boldsymbol{\Theta} - \boldsymbol{G}_s\boldsymbol{\eta}||^2 \\ &= ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2 - 2(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s))^T\boldsymbol{G}_s\boldsymbol{W}_s\boldsymbol{\Theta} \\ &+ \boldsymbol{W}_s^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{W}_s : \boldsymbol{\Theta}\boldsymbol{\Theta}^T \\ &- 2\boldsymbol{\eta}^T\boldsymbol{G}_s^T\left(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s) - \boldsymbol{G}_s\boldsymbol{W}_s\boldsymbol{\Theta}\right) \\ &+ \boldsymbol{\eta}^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{\eta}.\end{aligned} \tag{4.21}$$

We introduce a second *approximation* in terms of the family of q's over which we wish to optimize by using a *mean-field* decomposition (see Equation (2.19)) of the form:

$$\begin{aligned}q(\boldsymbol{\Theta},\boldsymbol{\eta},s,\tau) &\approx q(\boldsymbol{\Theta},\boldsymbol{\eta},s)\ q(\tau) \\ &= q(\boldsymbol{\Theta},\boldsymbol{\eta}|s)\ q(s)\ q(\tau) \\ &\approx q(\boldsymbol{\Theta}|s)\ q(\boldsymbol{\eta}|s)\ q(s)\ q(\tau).\end{aligned} \tag{4.22}$$

In the first line, τ is assumed to be a posteriori independent of the remaining latent variables on the premise that the measurement noise precision is determined by the experimental conditions and is not directly dependent on the latent variables. In the the third line, we assume that $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ are *conditionally* independent given s.[1] The latter assumption is justified by the role of $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ in the representation of $\boldsymbol{\Psi}$ (Equation (4.3)) expressing the main effects around the mean and the residual "noise" respectively. As such, it is also reasonable to assume that the means of $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ are zero a posteriori, i.e., $< \boldsymbol{\Theta} >_{q(\boldsymbol{\Theta}|s)} = \mathbf{0}$ and $< \boldsymbol{\eta} >_{q(\boldsymbol{\eta}|s)} = \mathbf{0}$. Furthermore, we employ an *isotropic* covariance for $\boldsymbol{\eta}$, i.e., $< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{q(\boldsymbol{\eta}|s)} = \lambda_{\eta,s}^{-1}\boldsymbol{I}_{d_\Psi}$ where $\lambda_{\eta,s}^{-1}$ represents the (unknown) variance.

If we denote the expectations with respect to $q(\tau)$, $q(\boldsymbol{\Theta}|s)$ and $q(\boldsymbol{\eta}|s)$ with $< . >_\tau$, $< . >_{\boldsymbol{\Theta}|s}$ and $< . >_{\boldsymbol{\eta}|s}$, then Equation (4.19) becomes [2]:

$$
\begin{aligned}
&\hat{\mathcal{F}}(q(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s),\boldsymbol{T}) \\
&= \tfrac{d_y}{2} < \log\tau >_\tau && (< \log p(\hat{\boldsymbol{y}}|s,\boldsymbol{\Theta},\boldsymbol{\eta},\tau,\boldsymbol{T}) >_q) \\
&- \tfrac{<\tau>_\tau}{2} \textstyle\sum_s q(s)\ ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2 \\
&+ < \tau >_\tau \textstyle\sum_s q(s)\ (\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s))^T \boldsymbol{G}_s\boldsymbol{W}_s < \boldsymbol{\Theta} >_{\boldsymbol{\Theta}|s} && (= 0\ since < \boldsymbol{\Theta} >_{\boldsymbol{\Theta}|s} = \mathbf{0}\ ) \\
&- \tfrac{<\tau>_\tau}{2} \textstyle\sum_s q(s)\ \boldsymbol{W}_s^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{W}_s :< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}|s} \\
&+ < \tau >_\tau \textstyle\sum_s q(s)\ < \boldsymbol{\eta} >_{\boldsymbol{\eta}|s}^T \boldsymbol{G}_s^T(\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)) && (= 0\ since < \boldsymbol{\eta} >_{\boldsymbol{\eta}|s} = \mathbf{0}\ ) \\
&- < \tau >_\tau \textstyle\sum_s q(s)\ < \boldsymbol{\eta} >_{\boldsymbol{\eta}|s}^T \boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{W}_s < \boldsymbol{\Theta} >_{\boldsymbol{\Theta}|s} && (= 0\ since < \boldsymbol{\eta} >_{\boldsymbol{\eta}|s} = \mathbf{0}\ ) \\
&- \tfrac{<\tau>_\tau}{2} \textstyle\sum_s q(s)\ \boldsymbol{G}_s^T\boldsymbol{G}_s :< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{\boldsymbol{\eta}|s} \\
&+ \textstyle\sum_s q(s)\ \log\frac{1}{S} && (< \log p_s(s) >_s) \\
&+ (a_0 - 1) < \log\tau >_\tau - b_0 < \tau >_\tau && (< \log p_\tau(\tau) >_\tau) \\
&+ \textstyle\sum_s q(s)\ (\frac{1}{2}\log|\boldsymbol{\Lambda}_{0,s}| - \frac{1}{2}\boldsymbol{\Lambda}_0 :< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}|s}) && (< \log p_\Theta(\boldsymbol{\Theta}|s) >_q) \\
&+ \textstyle\sum_s q(s)\ (\frac{d_\Psi}{2}\log\lambda_{0,\eta,s} - \frac{\lambda_{0,\eta,s}}{2}\boldsymbol{I} :< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{\boldsymbol{\eta}|s}) && (< \log p_\eta(\boldsymbol{\eta}|s) >_q) \\
&- \textstyle\sum_s q(s)\ \int q(\boldsymbol{\Theta}|s)\log q(\boldsymbol{\Theta}|s)\ d\boldsymbol{\Theta} && (- < \log q(\boldsymbol{\Theta}|s) >_q) \\
&- \textstyle\sum_s q(s)\ \int q(\boldsymbol{\eta}|s)\log q(\boldsymbol{\eta}|s)\ d\boldsymbol{\Theta} && (- < \log q(\boldsymbol{\eta}|s) >_q) \\
&- \textstyle\sum_s q(s)\ \log q(s) && (- < \log q(s) >_s) \\
&- < \log q(\tau) >_\tau . && (- < \log q(\tau) >_\tau)
\end{aligned}
\tag{4.23}
$$

Despite the long expression, the optimization of $\hat{\mathcal{F}}$ in the **VB-Expectation** step can be done *analytically* and we find that the optimal q (given $\boldsymbol{T}$) is:

$$
\begin{aligned}
&q^{opt}(\boldsymbol{\Theta}|s) \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_s^{-1}), \\
&q^{opt}(\boldsymbol{\eta}|s) \equiv \mathcal{N}(\mathbf{0}, \lambda_{\eta,s}^{-1}\boldsymbol{I}_{d_\Psi}), \\
&q^{opt}(\tau) \equiv Gamma(a,b),
\end{aligned}
\tag{4.24}
$$

where:

$$
\boldsymbol{\Lambda}_s = \boldsymbol{\Lambda}_{0,s} + < \tau >_\tau \boldsymbol{W}_s^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{W}_s, \tag{4.25}
$$

[1] This implies that $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$ are actually dependent, as one would expect.

[2] We omit constants that do not affect the optimization.

$$\lambda_{\eta,s} = \lambda_{0,\eta,s} + \frac{1}{d_\Psi} < \tau >_\tau tr(\boldsymbol{G}_s^T \boldsymbol{G}_s), \tag{4.26}$$

$$a = a_0 + d_y/2, \tag{4.27}$$

$$b = b_0 + \frac{1}{2}\sum_{s=1}^{S} q(s) \left(||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2 + \boldsymbol{W}_s^T \boldsymbol{G}_s^T \boldsymbol{G}_s \boldsymbol{W}_s : \boldsymbol{\Lambda}_s^{-1} + \lambda_{\eta,s}^{-1} \, tr(\boldsymbol{G}_s^T \boldsymbol{G}_s) \right). \tag{4.28}$$

Furthermore, for the latent variable s we find that:

$$q^{opt}(s) \propto e^{c_s}, \tag{4.29}$$

where:

$$c_s = \frac{1}{2} \log \frac{|\boldsymbol{\Lambda}_{0,s}|}{|\boldsymbol{\Lambda}_s|} + \frac{d_\Psi}{2} \log \frac{\lambda_{0,\eta,s}}{\lambda_{\eta,s}} - \frac{< \tau >_\tau}{2} ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2, \tag{4.30}$$

and $< \tau >_\tau = \frac{a}{b}$. The normalization constant for $q(s)$ can be readily found by imposing the condition $\sum_{s=1}^{S} q^{opt}(s) = 1$ which yields:

$$q^{opt}(s) = \frac{e^{c_s}}{\sum_{s'} e^{c_{s'}}}. \tag{4.31}$$

While the optimal q's are inter-dependent, we note in the expression above that the posterior probability of each mixture component s, as one would expect, increases as the mean-square error $||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2$ gets smaller. More interestingly perhaps, $q^{opt}(s)$ increases as the determinant of the posterior precision matrix $\boldsymbol{\Lambda}_s$ decreases, i.e., as the posterior variance associated with the reduced coordinates $\boldsymbol{\Theta}$ of component s increases. The same effect is observed for the posterior residual variance $\lambda_{\eta,s}^{-1}$. This implies that, ceteris paribus, mixture components with larger posterior variance will have a bigger weight in the overall posterior.

For the optimal q^{opt} (given $\boldsymbol{T}$) in the equations above, the variational lower bound $\hat{\mathcal{F}}$ takes the following form (terms independent of q^{opt} or $\boldsymbol{T}$ are omitted - for details see Appendix B):

$$\begin{aligned} \hat{\mathcal{F}}(q^{opt}, \boldsymbol{T}) \;&= \sum_{s=1}^{S} q^{opt}(s) \left(-\tfrac{<\tau>_\tau}{2} ||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2 + \tfrac{1}{2} \log \tfrac{|\boldsymbol{\Lambda}_{0,s}|}{|\boldsymbol{\Lambda}_s|} + \tfrac{d_\Psi}{2} \log \tfrac{\lambda_{0,\eta,s}}{\lambda_{\eta,s}} \right) \\ &\quad - \sum_{s=1}^{S} q^{opt}(s) \; \log q^{opt}(s) \\ &\quad + a \log(< \tau >_\tau), \end{aligned} \tag{4.32}$$

and

$$\mathcal{F}(q^{opt}, \boldsymbol{T}) = \hat{\mathcal{F}}(q^{opt}, \boldsymbol{T}) + \log p_T(\boldsymbol{T}), \tag{4.33}$$

where $Z(a,b) = \frac{\Gamma(a)}{b^a}$ is the normalization constant of a $Gamma$ distribution with parameters a, b. This can be computed at each full iteration of VB-EM in order to monitor convergence.

While it is difficult again to gain insight in the expression above due to the inter-dependencies between the various terms, we note that the smaller the mean-square error of $||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_s)||^2$ becomes (i.e., the better the mean $\boldsymbol{\mu}_s$ is able to reproduce the measurements), the more the lower bound increases. In addition we can see that the lower bound increases as the variance of the mixture components $\boldsymbol{\Lambda}_s^{-1}, \lambda_{\eta,s}^{-1}$ gets larger, meaning the more variance they capture.

For the **VB-Maximization** step, it can be readily established from Equation (4.23) that the optimization of $\mathcal{F}$ with respect to $\boldsymbol{\mu}$ (given q) involves the following set of *uncoupled* optimization problems:

$$\max_{\boldsymbol{\mu}_j} \mathcal{F}_{\mu_j} = -\frac{<\tau>_\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}(\boldsymbol{\mu}_j)||^2 + \log p_\mu(\boldsymbol{\mu}_j), \quad j = 1, \ldots, S. \tag{4.34}$$

Since the objectives are identical for each j, we can deduce that $\boldsymbol{\mu}_j$ should correspond to (different or identical) local maxima of $\mathcal{F}$. This implies that in the posterior approximation constructed, each Gaussian in the mixture is associated with a (regularized - due to the prior) local optimum in the least-square solution of the inverse problem. The search for multiple local optima, and more importantly their number, is discussed in the next section.

The determination of the optimal $\boldsymbol{\mu}_j$ is performed using first-order derivatives of $\frac{\partial \mathcal{F}_{\mu_j}}{\partial \boldsymbol{\mu}_j}$. Since $\log p_\mu(\boldsymbol{\mu}_j)$ and its derivative $\frac{\partial \log p_\mu(\boldsymbol{\mu}_j)}{\partial \boldsymbol{\mu}_j}$ are analytically unavailable, we employ an additional layer (inner loop) of Expectation-Maximization to deal with the hyperparameters in the prior of $\boldsymbol{\mu}_j$. The details were discussed in Section 3.2.4 and Appendix A for a single mixture component which is now applied for all mixture components j separately.

Considering the *computational cost* of these operations, we point out that the updates of $\boldsymbol{\mu}_j$ are the most demanding as they require calls to the forward model to evaluate $\boldsymbol{y}(\boldsymbol{\mu}_j)$ and the derivatives $\boldsymbol{G}_j = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}_j}|_{\boldsymbol{\Psi}_j = \boldsymbol{\mu}_j}$, details in Appendix D. For the computation of the derivatives $\boldsymbol{G}_j$ we employ the adjoint formulations which offer great savings when $d_\Psi \gg d_y$ [155]. As discussed in detail in Section 2.3.4, the latter condition can be removed as long as a direct solver is used for the solution of the forward problem. In this case, the cost of the solution of the adjoint equations is even less than that of the forward solution.

The remaining aspect of the **VB-Maximization** step involves the optimization with respect to the $\boldsymbol{W}$ (given q). As with $\boldsymbol{\mu}$, it suffices to consider only the terms in Equation (4.23) that depend on $\boldsymbol{W}$ (which we denote by $\mathcal{F}_{W_j}$) and which again lead to a set of S uncoupled problems:

$$\max_{\boldsymbol{W}_j} \mathcal{F}_{W_j} = -\frac{<\tau>_\tau}{2}(\boldsymbol{W}_j^T \boldsymbol{G}_j^T \boldsymbol{G}_j \boldsymbol{W}_j) : \boldsymbol{\Lambda}_j^{-1} + \log p_W(\boldsymbol{W}_j), \quad j = 1, \ldots, S. \tag{4.35}$$

The first term prefers directions corresponding to the smallest eigenvectors of $\boldsymbol{G}_j^T \boldsymbol{G}_j$, where $\boldsymbol{G}_j = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}_j}|_{\boldsymbol{\Psi}_j = \boldsymbol{\mu}_j}$ is the gradient of the map at $\boldsymbol{\mu}_j$. As discussed previously in Section 4.2.2, the prior $p_W(\boldsymbol{W}_j)$ enforces the orthogonality of the basis vectors in $\boldsymbol{W}_j$. To solve this constrained optimization problem, we use the iterative algorithm of [184], which employs a Cayley transform to enforce the constraint. It makes use of first-order derivatives of $\mathcal{F}_{W_j}$ and as such does not require any additional forward model runs.

With regard to the number of columns d_Θ in $\boldsymbol{W}_j$ (which is equal to the dimension of $\boldsymbol{\Theta}$), we assume that this is the same across all mixture components S. We had developed an information-theoretic criterion in Section 3.2.5 which can also be employed here. This allows the adaptive determination of d_Θ by measuring the information gain, here denoted by $I(d_\Theta, j)$ for each mixture component j, that each new dimension in $\boldsymbol{\Theta}$ furnishes. When these fall below a threshold I_{max} (in our examples we use $I_{max} = 1\%$), i.e.:

$$I(d_\Theta) = \max_j I(d_\Theta, j) \leq I_{max}, \tag{4.36}$$

we assume that the number of $\boldsymbol{\Theta}$ is sufficient. A detailed discussion on the estimation of d_Θ using the information gain $I(d_\Theta, j)$ is given in Section 3.2.5 which is extended to the multimodal case in Appendix C.

Following the previous discussion in Equation (4.6), we note that once the (approximate) posterior $q(\boldsymbol{\Theta}, \boldsymbol{\eta}, \tau, s)$ and the optimal model parameters $\boldsymbol{T}$ have been computed, we obtain a *multimodal* posterior approximation for the material parameters $\boldsymbol{\Psi}$, which is given by:

$$\begin{aligned} p(\boldsymbol{\Psi}|\boldsymbol{T}, \hat{\boldsymbol{y}}) &= \textstyle\sum_{s=1}^S \int \delta(\boldsymbol{\Psi} - (\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}))\ p(s, \boldsymbol{\Theta}, \boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{W}, \hat{\boldsymbol{y}})\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta} \\ &\approx \textstyle\sum_{s=1}^S \int \delta(\boldsymbol{\Psi} - (\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}))\ q(s, \boldsymbol{\Theta}, \boldsymbol{\eta})\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta} \\ &= \textstyle\sum_{s=1}^S q(s) \int \delta(\boldsymbol{\Psi} - (\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}))\ q(\boldsymbol{\Theta}, \boldsymbol{\eta}|s)\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta} \\ &= \textstyle\sum_{s=1}^S q(s)\ q_s(\boldsymbol{\Psi}) = q(\boldsymbol{\Psi}), \end{aligned} \tag{4.37}$$

where each component in the last mixture is given by:

$$\begin{aligned} q_s(\boldsymbol{\Psi}) &= \int \delta(\boldsymbol{\Psi} - (\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta} + \boldsymbol{\eta}))\ q(\boldsymbol{\Theta}, \boldsymbol{\eta}|s)\ d\boldsymbol{\Theta}\ d\boldsymbol{\eta} \\ &\equiv \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{D}_s), \end{aligned} \tag{4.38}$$

i.e., a multivariate Gaussian with mean $\boldsymbol{\mu}_s$ and covariance $\boldsymbol{D}_s$ where:

$$\boldsymbol{D}_s = \boldsymbol{W}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{W}_s^T + \lambda_{\eta,s}^{-1} \boldsymbol{I}_{d_\Psi}. \tag{4.39}$$

Based on Equation (4.37), one can evaluate the posterior mean and covariance of

$\boldsymbol{\Psi}$ as follows:

$$\begin{aligned}
< \boldsymbol{\Psi} >_q &= << \boldsymbol{\Psi}|s >_q >= \langle \boldsymbol{\mu}_s \rangle = \textstyle\sum_{s=1}^{S} q(s)\boldsymbol{\mu}_s, \\
\mathrm{Cov}_q[\boldsymbol{\Psi}] &= < \boldsymbol{\Psi}\boldsymbol{\Psi}^T >_q - < \boldsymbol{\Psi} >_q < \boldsymbol{\Psi} >_q^T = << \boldsymbol{\Psi}\boldsymbol{\Psi}^T|s >_q > - < \boldsymbol{\Psi}_q >< \boldsymbol{\Psi} >_q^T \\
&= \textstyle\sum_{s=1}^{S} q(s)(\boldsymbol{D}_s + \boldsymbol{\mu}_s\boldsymbol{\mu}_s^T) - \left(\sum_{s=1}^{S} q(s)\boldsymbol{\mu}_s\right)\left(\sum_{s=1}^{S} q(s)\boldsymbol{\mu}_s\right)^T \\
&= \textstyle\sum_{s=1}^{S} q(s)\boldsymbol{D}_s + \sum_{s=1}^{S} q(s)\boldsymbol{\mu}_s\boldsymbol{\mu}_s^T - \left(\sum_{s=1}^{S} q(s)\boldsymbol{\mu}_s\right)\left(\sum_{s=1}^{S} q(s)\boldsymbol{\mu}_s^T\right).
\end{aligned} \tag{4.40}$$

Posterior moments of any order or posterior probabilities can be readily computed as well.

Note that if $\boldsymbol{\Lambda}_s^{-1}$ is diagonalized, e.g., $\boldsymbol{\Lambda}_s^{-1} = \boldsymbol{U}_s\hat{\boldsymbol{\Lambda}}_s^{-1}\boldsymbol{U}_s^T$ where $\hat{\boldsymbol{\Lambda}}_s^{-1}$ is diagonal and $\boldsymbol{U}_s$ contains the eigenvectors of $\boldsymbol{\Lambda}_s^{-1}$, then:

$$\begin{aligned}
\boldsymbol{D}_s &= \boldsymbol{W}_s\boldsymbol{U}_s\hat{\boldsymbol{\Lambda}}_s^{-1}\boldsymbol{U}_s^T\boldsymbol{W}_s^T + \lambda_{\eta,s}^{-1}\boldsymbol{I}_{d_\Psi} \\
&= \hat{\boldsymbol{W}}_s\hat{\boldsymbol{\Lambda}}_s^{-1}\hat{\boldsymbol{W}}_s^T + \lambda_{\eta,s}^{-1}\boldsymbol{I}_{d_\Psi}.
\end{aligned} \tag{4.41}$$

Each $\hat{\boldsymbol{W}}_s$ is also orthogonal and contains the d_Θ principal directions of posterior covariance of $\boldsymbol{\Psi}_s$. Therefore, we see that in the VB-E step it suffices to consider an approximate posterior $q(\boldsymbol{\Theta}|s)$ with a diagonal covariance, e.g., $\boldsymbol{\Lambda}_s = diag(\lambda_{s,i}), i = 1, ..., d_\Theta$. As a consequence the update equation for $\boldsymbol{\Lambda}_s$ (Equation (4.25)) reduces to:

$$\lambda_{s,i} = \lambda_{0,s,i} + < \tau >_\tau \boldsymbol{w}_{s,i}^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{w}_{s,i}, \tag{4.42}$$

where $\boldsymbol{w}_{s,i}$ is the i^{th} column vector of $\boldsymbol{W}_s$.

We note that in all the aforementioned expressions we assumed that the number of components S is given and fixed. Nevertheless, if for some s, $q^{opt}(s)$ is zero (or negligible), the corresponding component will have no (posterior) contribution. In Algorithm 2 we summarize the main steps of the algorithm for a fixed S. Steps $5-7$ correspond to the aforementioned VB-Expectation and steps 2 and 4 to the VB-Maximization step. In the next section we discuss an adaptive strategy for determining S.

4.2.4 Finding the required number of mixture components S

A critical component of the proposed framework is the cardinality S of the model, i.e., the number of modes in the approximation of the posterior. The mean $\boldsymbol{\mu}_j$ of each Gaussian component is optimal when it corresponds to a local maximum of the objective in Equation (4.34), but suboptimal solutions can be found by using suboptimal $\boldsymbol{\mu}_j$.

A consistent way of carrying out this model selection task, within the advocated Bayesian framework, is to compute or approximate the model evidence term $p(\hat{\boldsymbol{y}})$ in Equation (4.14) for various values of S. This can be followed by selecting the one that

Algorithm 2 Algorithm for fixed S

while $\mathcal{F}_\mu$ in Equation (4.34), has not converged **do**
  For $j = 1 : S$: Optimize $\boldsymbol{\mu}_j$ using Equation (4.34)
  while $\mathcal{F}$ in Equation (4.33), has not converged **do**
    For $j = 1 : S$: Optimize $\boldsymbol{W}_j$ using Equation (4.35)
    For $s = 1 : S$: Update $q(\boldsymbol{\Theta}|s) \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_s^{-1})$ using Equation (4.42)
    For $s = 1 : S$: Update $q(\boldsymbol{\eta}|s) \equiv \mathcal{N}(\mathbf{0}, \lambda_{\eta,s}^{-1}\boldsymbol{I}_{d_\Psi})$ using Equation (4.26)
    Update $q(\tau) \equiv Gamma(a, b)$ and $q(s)$ using Equations (4.27, 4.28, 4.31)
  end while
end while

gives the largest $p(\hat{\boldsymbol{y}})$ or performing model averaging with probabilities proportional to these terms for each values of S [54, 79]. Nevertheless computing $p(\hat{\boldsymbol{y}})$ is impractical as it requires integrating over all parameters including the high-dimensional $\boldsymbol{T}$, i.e., a fully Bayesian treatment of the $\boldsymbol{\mu}$ and $\boldsymbol{W}$. In the formulation presented thus far however, we computed point estimates by maximizing the variational bound $\mathcal{F}$ to the log posterior $p(\boldsymbol{T}|\hat{\boldsymbol{y}})$ (Equation (4.16)).

One might be inclined to compute this $\mathcal{F}$ (assuming it is a good approximation of $p(\boldsymbol{T}|\hat{\boldsymbol{y}})$) for different values of S and use it to identify the optimal S. We note though that such terms are not comparable as they depend on the number of parameters in $\boldsymbol{T}$ which changes with S. As a result, such comparisons would be meaningless. As a third option one could potentially employ one of the well-known approximate validation metrics, e.g., AIC or BIC, which penalize the log posterior ($p(\boldsymbol{T}|\hat{\boldsymbol{y}})$ or $\mathcal{F}$) with the number of parameters, but these are known to be valid only in limiting cases, for large datasets [79, 207].

Furthermore, we note that if two components ($S = 2$) with the same $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (and as a result $\boldsymbol{G}_1 = \boldsymbol{G}_2$, and $\boldsymbol{W}_1 = \boldsymbol{W}_2$, $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$) are considered, then $q(s = 1) = q(s = 2) = \frac{1}{2}$. Even though a mixture of these two identical components gives rise to a single Gaussian (Equation (4.38)), it is obvious that the second component provides no new information regarding the posterior. This is because the posterior $p(s|\hat{\boldsymbol{y}})$ (and its approximation $q(s)$) accounts for the *relative* plausibility (as compared to the other components) that the component s could have given rise to a $\boldsymbol{\Psi}$ (that in turn gave rise to $\boldsymbol{y}(\boldsymbol{\Psi})$) that matches the observations $\hat{\boldsymbol{y}}$.

For this purpose, we advocate an adaptive algorithm (Algorithm 3) that proposes new components (component birth) and removes those (component death) that do not furnish new information.

Algorithm 3 Adaptive algorithm for the determination of appropriate S

1: Initialize $S = S_0$ (e.g., $S_0 = 1$), $L = 0$, $iter = 0$. Set prior hyperparameters $a_0, b_0, \{\lambda_{0,j}, \lambda_{0,\eta,j}\}_{j=1}^{S}$. Initialize $\{\boldsymbol{\mu}_j, \boldsymbol{W}_j\}_{j=1}^{S}$ randomly (as long as $\boldsymbol{W}_j$ are orthogonal) and call Algorithm 2.
2: **while** $L < L_{max}$ **do**
3: $iter \leftarrow iter + 1$
4: (Component Birth) Propose ΔS new mixture components and initialize $\boldsymbol{\mu}_j$ for $j = S+1, \ldots, S+\Delta S$ according to Equation (4.43).
5: Call Algorithm 2
6: (Component Death) Delete any of the new components that satisfy the component death criterion in Equation (4.44)
7: Compute $q(s)$ of surviving components (Equation (4.31)), remove any components with $q(s) < q_{min}$ [3] and update S
8: **if** None of the ΔS new components remain active **then**
9: $L \leftarrow L + 1$;
10: **else**
11: $L \leftarrow 0$;
12: **end if**
13: **end while**

We discuss in detail the steps above that contain new features as compared to Algorithm 2:

- Steps 2 and 8-12:
  The overall algorithm is terminated when L_{max} successive attempts to add new mixture components have failed (in all examples discussed $L_{max} = 3$). L counts the number of successive failed attempts to add new components and $iter$ the total number of component birth attempts. During each of those, ΔS new mixture components are proposed (component birth) and optimized. Since the $\boldsymbol{\mu}$-updates of each mixture component imply a certain number of forward model solutions, the termination criterion could be alternatively expressed in terms of the maximum allowable number of such forward calls.[4]

- Step 4 (Component Birth):
  Given S mixture components, we propose the addition of ΔS new components. Their means $\boldsymbol{\mu}_{j_{new}}$, for $j_{new} = S+1, \ldots, S+\Delta S$, are initialized by perturbing the mean of one of the pre-existing S components as follows: We pick the mixture component $j_{parent} \in 1, \ldots, S$ that has the smallest contribution in the lower

[3] Throughout this work we use $q_{min} = 1 \times 10^{-3}$.

[4] See Figure 4.6 where for $iter = 2$, none of the $\Delta S = 3$ mixture components survive. Here L increases from 0 to 1.

bound $\hat{\mathcal{F}}$ in Equation (4.33), and therefore provides the worst fit to the data.[5] We initialize $\boldsymbol{\mu}_{j_{new}}$ randomly as follows:

$$\boldsymbol{\mu}_{j_{new}} = \boldsymbol{\mu}_{j_{parent}} + \boldsymbol{W}_{j_{parent}}\boldsymbol{\Theta} + \alpha\boldsymbol{\eta}, \tag{4.43}$$

where $\boldsymbol{\Theta}$ is sampled from the posterior $q(\boldsymbol{\Theta}|s = j_{parent})$ and $\boldsymbol{\eta}$ is sampled from $q(\boldsymbol{\eta}|s = j_{parent})$. The role of α is to amplify the perturbations. The value of $\alpha = 10$ was used throughout this work. Very large α increase the possibility of finding a new mode but increase the number of $\boldsymbol{\mu}$-updates and therefore the number of forward model calls. The remaining model parameters for each new component are initialized as usual and are updated according to the VB-EM scheme discussed in Step 5.

- Step 5:
  Whereas the VB-EM scheme discussed in the previous section has to be run every time when new components are proposed (i.e., S changes), we note here that the updates for the pre-existing components require only very few new (if any) forward-model runs. This is because updates for pre-existing $\boldsymbol{\mu}_j$ (Equation (4.34)) are only required if $< \tau >_\tau$ changes. While $< \tau >_\tau$ is affected by all components S (old and new, Equation (4.28)), it generally does not change significantly after the first few components.

- Step 6 (Component Death):
  We employ an information-theoretic criterion that measures the discrepancy ('similarity distance') $d_{j_{old},j_{new}}$ between a new component $j_{new} \in \{S+1, \ldots, S+\Delta S\}$ and an existing one $j_{old} \in \{1, \ldots, S\}$. If this is smaller than a prescribed threshold d_{min}, *for any of the existing components* j_{old}, then the component j_{new} is removed as the two mixture components are too close to each other. In other words, the component death criterion may be stated as:

$$\text{if } \exists j_{old} \text{ such that } d_{j_{old},j_{new}} < d_{min}. \tag{4.44}$$

  Throughout this work, we use $d_{min} = 0.01$[6] and define $d_{j_{old},j_{new}}$ as follows:

$$d_{j_{old},j_{new}} = \frac{KL(q_{j_{old}}||q_{j_{new}})}{d_\Psi}, \tag{4.45}$$

where the KL divergence between two multivariate Gaussians $q_{j_{old}}(\boldsymbol{\Psi})$ and

[5] If a specific mixture component has already been used as a parent in a previous unsuccessful attempt, the next worst mixture component is used.

[6] Nevertheless, as shown in the numerical examples, a much lower value of $d_{min} = 10^{-8}$ would have yielded identical results.

$q_{j_{new}}(\boldsymbol{\Psi})$ (Equation (4.38)) can be analytically computed as:

$$\begin{aligned} KL(q_{j_{old}}(\boldsymbol{\Psi})||q_{j_{new}}(\boldsymbol{\Psi})) = \; & \tfrac{1}{2}\log|\boldsymbol{D}_{j_{new}}| + \tfrac{1}{2}\boldsymbol{D}_{j_{new}}^{-1} : \boldsymbol{D}_{j_{old}} \\ & + \tfrac{1}{2}(\boldsymbol{\mu}_{j_{old}} - \boldsymbol{\mu}_{j_{new}})^T \boldsymbol{D}_{j_{new}}^{-1}(\boldsymbol{\mu}_{j_{old}} - \boldsymbol{\mu}_{j_{new}}) \\ & - \tfrac{1}{2}\log|\boldsymbol{D}_{j_{old}}| - \tfrac{d_\Psi}{2}. \end{aligned} \tag{4.46}$$

We note that such a discrepancy metric takes the whole distribution into account and not just the locations of the modes $\boldsymbol{\mu}_{j_{old}}$, $\boldsymbol{\mu}_{j_{new}}$.The denominator d_Ψ of the KL divergence normalizes it with respect to the number of dimensions. Therefore, $d_{j_{old},j_{new}}$ is the average KL divergence over the d_Ψ dimensions and d_{min} expresses the minimum acceptable averaged KL distance per dimension.

In Figure 4.4, we plot for illustration purposes the contour lines of the KL divergence of various one-dimensional Gaussians $\mathcal{N}(\mu, \sigma^2)$ (as a function of their mean μ and variance σ^2) with respect to the standard Gaussian $\mathcal{N}(0, 1)$. We note that other distance metrics, e.g., the Fisher information matrix, could have equally been used.

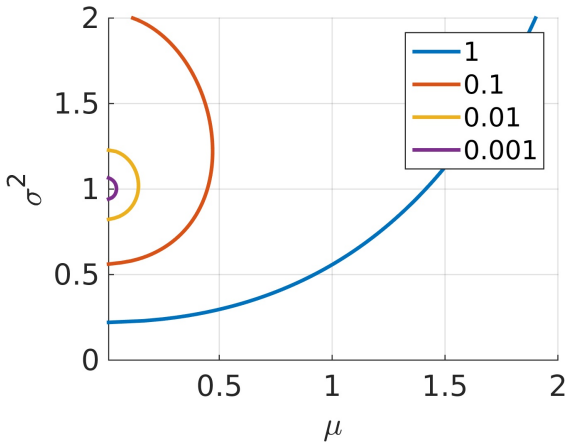


Figure 4.4: Contour lines of the KL-divergence between $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, \sigma^2)$ with respect to μ, σ^2. Any Gaussian with (μ, σ^2) within the yellow line would be deleted according to the criterion defined.

We can take advantage of the low-rank decomposition $\boldsymbol{D}_s$ in Equation (4.39) in order to efficiently compute the inverse of $\boldsymbol{D}_{j_{new}}^{-1}$ as:

$$\begin{aligned} \boldsymbol{D}_s^{-1} &= (\boldsymbol{W}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{W}_s^T + \lambda_{\eta,s}^{-1} \boldsymbol{I}_{d_\Psi})^{-1} \\ &= \lambda_{\eta,s} \boldsymbol{I}_{d_\Psi} - \lambda_{\eta,s}^2 \boldsymbol{W}_s \underbrace{(\boldsymbol{\Lambda}_s + \lambda_{\eta,s} \boldsymbol{I})^{-1}}_{diagonal} \boldsymbol{W}_s^T. \end{aligned} \tag{4.47}$$

Similarly, the determinants can be readily computed as:

$$\begin{aligned} |\boldsymbol{D}_s| &= |\boldsymbol{W}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{W}_s^T + \lambda_{\eta,s}^{-1} \boldsymbol{I}_{d_\Psi}| \\ &= |\boldsymbol{\Lambda}_s + \lambda_{\eta,s} \boldsymbol{I}| \; |\boldsymbol{\Lambda}_s^{-1}| \; \lambda_{\eta,s}^{-d_\Psi}. \end{aligned} \tag{4.48}$$

4.2.5 Verification - Combining VB approximations with importance sampling

The framework advocated is based on two approximations: a) linearization of the response (Equation (4.20)) and, b) the mean-field decomposition of the approximating distribution (Equation (4.22)). This unavoidably introduces bias and the approximate posterior will deviate from the exact. In order to assess the quality of the approximation but also to correct for any bias in the posterior estimates, we propose using importance sampling (IS) (Section 2.2.3). In particular, we employ the approximate conditional posterior q as the importance sampling density and compute the effective sample size (ESS).

The performance of IS can decay rapidly in high dimensions [79] and due to the fact that $\boldsymbol{\eta}$ has a negligible effect in the inferred posterior, we propose using $p(\boldsymbol{\Theta}, s|\hat{\boldsymbol{y}}, \boldsymbol{T})$ as the target density. According to Equation (4.15):

$$\begin{aligned} p(\boldsymbol{\Theta}, s|\hat{\boldsymbol{y}}, \boldsymbol{T}) &= \int p(\boldsymbol{\Theta}, s, \tau|\hat{\boldsymbol{y}}, \boldsymbol{T})\ d\tau \\ &\propto \int p(\hat{\boldsymbol{y}}|s, \boldsymbol{\Theta}, \tau, \boldsymbol{T})\ p_\tau(\tau)\ p_\Theta(\boldsymbol{\Theta}|s)\ p_s(s)\ d\tau \\ &\propto \int \tau^{d_y/2} e^{-\frac{\tau}{2}||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\mu}_s+\boldsymbol{W}_s\boldsymbol{\Theta})||^2}\ p_\tau(\tau)\ d\tau\ p_\Theta(\boldsymbol{\Theta}|s)\ p_s(s) \\ &= \frac{\Gamma(a_0+d_y/2)}{(b_0+\frac{||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\mu}_s+\boldsymbol{W}_s\boldsymbol{\Theta})||^2}{2})^{a_0+d_y/2}}\ p_\Theta(\boldsymbol{\Theta}|s)\ p_s(s), \end{aligned} \tag{4.49}$$

where the Gamma prior $p_\tau(\tau)$ is from Equation (4.8) and MAP estimates of $\boldsymbol{\mu}, \boldsymbol{W}$ are used. In cases where non-conjugate priors for τ are employed, the IS procedure detailed here has to be performed in the joint space $(\boldsymbol{\Theta}, s, \tau)$.

Given M samples $(\boldsymbol{\Theta}^{(m)}, s^{(m)})$ drawn from the mixture of Gaussians $q(\boldsymbol{\Theta}, s)$ in Equation (4.24) and Equation (4.31), IS reduces to computing the unnormalized weights $w^{(m)}$ as follows:

$$w^{(m)} = \frac{p(\boldsymbol{\Theta}^{(m)}, s^{(m)}|\hat{\boldsymbol{y}}, \boldsymbol{T})}{q(\boldsymbol{\Theta}^{(m)}, s^{(m)})}. \tag{4.50}$$

With $w^{(m)}$ (asymptotically) unbiased estimates, the expectations of any integrable function $g(\boldsymbol{\Psi})$ with respect to the exact posterior can be computed as:

$$\begin{aligned} < g(\boldsymbol{\Psi}) >_{p(\boldsymbol{\Theta}, s|\hat{\boldsymbol{y}}, \boldsymbol{\mu}, \boldsymbol{W})} &= \sum_{s=1}^{S} \int g(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta})\ p(\boldsymbol{\Theta}, s|\hat{\boldsymbol{y}}, \boldsymbol{T})\ d\boldsymbol{\Theta} \\ &= \sum_{s=1}^{S} \int g(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta})\ \frac{p(\boldsymbol{\Theta}, s|\hat{\boldsymbol{y}}, \boldsymbol{T})}{q(\boldsymbol{\Theta}, s)}\ q(\boldsymbol{\Theta}, s)\ d\boldsymbol{\Theta} \\ &= \sum_{m=1}^{M} \hat{w}^{(m)} g(\boldsymbol{\mu}_s + \boldsymbol{W}_s\boldsymbol{\Theta}^{(m)}), \end{aligned} \tag{4.51}$$

where the $\hat{w}^{(m)}$ are the normalized IS weights ($\sum_{m=1}^{M} \hat{w}^{(m)} = 1$):

$$\hat{w}^{(m)} = \frac{w^{(m)}}{\sum_{m'=1}^{M} w^{(m')}}. \tag{4.52}$$

In the following examples we employ estimators such as these to compute the asymptotically (as $M \to \infty$) *exact* posterior mean (i.e., $g(\mathbf{\Psi}) = \mathbf{\Psi}$), posterior variances as well as posterior quantiles. Furthermore in order to assess the overall accuracy of the approximation and to provide a measure of comparison with other inference strategies (past and future), we report the (normalized) effective sample size (ESS), according to Equation (2.11).

Finally, we note that if there are additional modes in the exact posterior that have not been discovered by $q(\mathbf{\Theta}, s)$, the ESS could still be misleadingly large (for large but finite sample sizes M). This however is a general problem of Monte Carlo-based techniques, i.e., they cannot reveal (unless $M \to \infty$) the presence of modes in the target density unless these modes are visited by samples.

4.3 Numerical illustration

We consider two numerical illustrations. The primary goal of the first example is to provide insight into the adaptive search algorithm for determining S and for that reason we analyze an one-dimensional, multimodal density. The second example pertains to the motivating application of elastography. We demonstrate how the proposed framework can reveal the presence of multiple modes and, when justified, can identify low-dimensional approximations for each of these modes with a limited number of forward calls. An overview of the most important quantities/dimensions of the following two examples is contained in Table 4.1.

	Example 1	Example 2
Dimension of observables: d_y	1	5100
Dimension of latent variables: d_Ψ	1	2500
Dimension of reduced latent variables: d_Θ	1	11
No. of forward calls	< 200	< 1200

Table 4.1: Summary of the number of observables, forward calls and the dimensionality reduction in the following two examples.

Toy Example

Our goal in this first example is solely to illustrate the features and capabilities of the adaptive search algorithm for determining the number of mixture components S. For that purpose we selected an one-dimensional example (in order to remove any effects from the dimensionality reduction) that can be semi-analytically investigated

and exhibits a multimodal posterior. We assume that the model equation is of the form:

$$y(\Psi) = \Psi^3 + \Psi^2 - \Psi, \quad \Psi \in \mathbb{R}, \tag{4.53}$$

and is depicted in Figure 4.5. Let $\Psi_{exact} = 0.8$ be the reference solution for which $y(\Psi_{exact}) = 0.352$. With the addition of noise it is assumed that the actual measurement is $\hat{y} = 0.45$. This is shown with a horizontal line in Figure 4.5, where for $\hat{y} = 0.45$ three modes for Ψ exist. The Gaussian prior on Ψ has zero mean and a variance of $\lambda_0 = 1 \times 10^{-10}$.

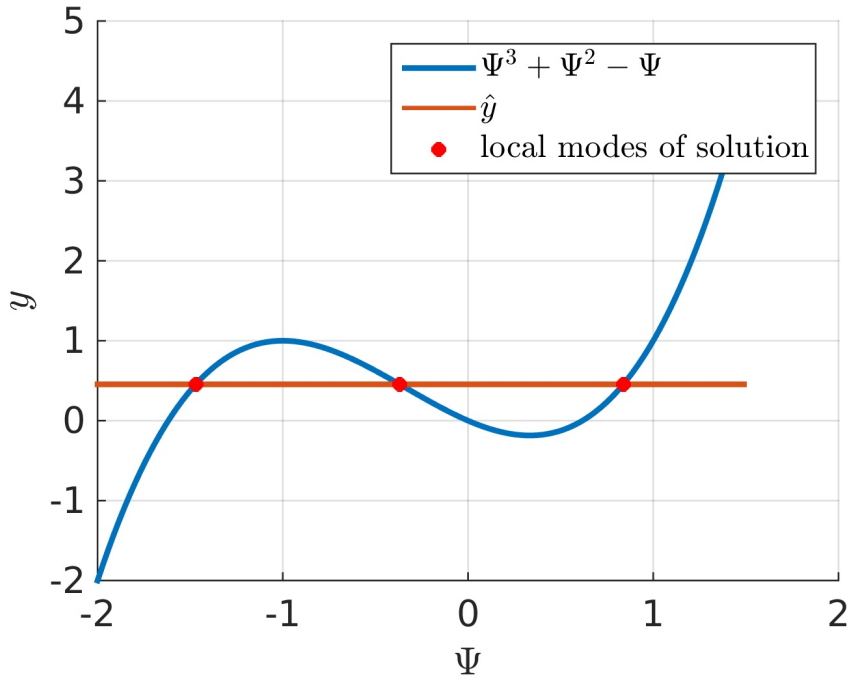


Figure 4.5: Polynomial $y = \Psi^3 + \Psi^2 - \Psi$. It can be seen that for the measurement at $\hat{y} = 0.45$ three possible solutions exist.

As this is an one-dimensional example, the dimensionality reduction aspects are invalid and $\boldsymbol{\eta}$ (Equation (4.3)) is also unnecessary. We initialize the adaptive Algorithm 3 with $S_0 = 4$ and propose/add $\Delta S = 3$ components at each iteration $iter$. We summarize the results produced by successive iterations in Figure 4.6. Two mixture components are identified at initialization (out of the $S_0 = 4$ proposed). Proposed components at subsequent iterations that do not survive are marked with a red cross.

Table 4.2 contains the values of the normalized KL-based discrepancy metric (Equation (4.45)) for all pairs of the 6 mixture components at $iter = 2$ (Figure 4.6). As it can be seen by the values, components 4, 5 and 6 satisfy the component death criterion (Equation (4.44)) and are therefore removed.

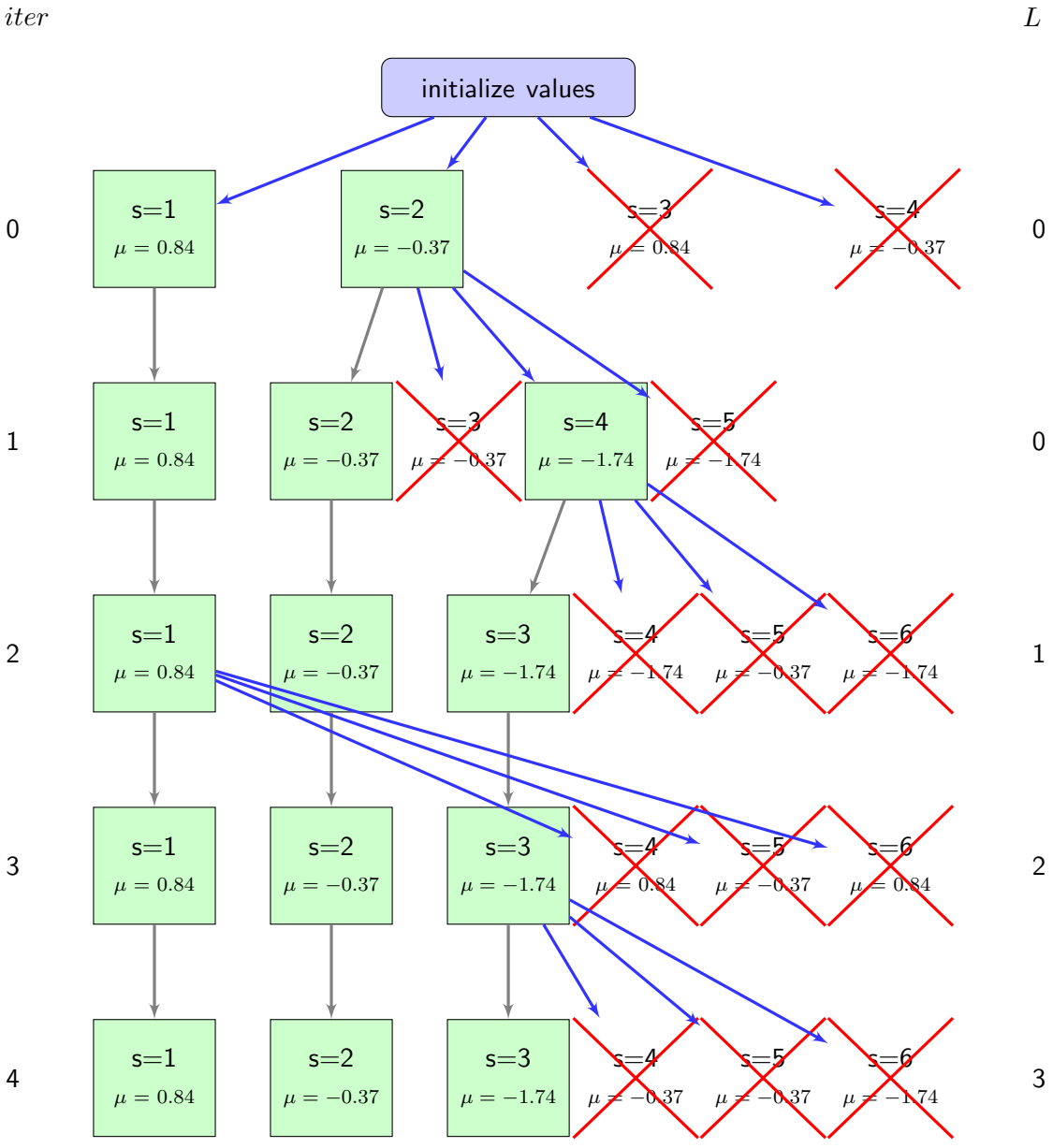


Figure 4.6: Evolution of Algorithm 3 for Example 1 with $S_0 = 4$ and $\Delta S = 3$. Green boxes correspond to surviving mixture components, whereas the ones that are deleted are marked with a red cross. The rows are numbered based on $iter$ and the value of L is reported on the right. The mean μ_j of each component is also reported in each box. Mixture components connected with gray arrows stay active whereas mixture components with blue arrows represent new initialized and updated mixture components which are then deactivated.

	s=1	s=2	s=3	s=4	s=5	s=6
s=1	0	61.97	824	824	61.97	824
s=2		0	188.4	188.4	$\mathbf{1.2 \times 10^{-10}}$	188.4
s=3			0	$\mathbf{2.3 \times 10^{-09}}$	51.59	$\mathbf{2.3 \times 10^{-09}}$
s=4				0	51.59	$\mathbf{2.3 \times 10^{-09}}$
s=5					0	188.4

Table 4.2: Normalized KL divergence (Equation (4.45)) between each pair of mixture components. Pairs which are very similar (see also the means in Figure 4.6) have a very small KL divergence (shown in bold).

The three components that persist have the following posterior probabilities:

$$q(s=1) = 0.24, \quad q(s=2) = 0.50, \quad q(s=3) = 0.26. \tag{4.54}$$

The Gaussians (Equation (4.37)) associated with each component are:

$$\begin{aligned} q(\Psi|s=1) &= \mathcal{N}(0.84, 0.00135), \\ q(\Psi|s=2) &= \mathcal{N}(-0.37, 0.00590), \\ q(\Psi|s=3) &= \mathcal{N}(-1.74, 0.00162). \end{aligned} \tag{4.55}$$

The algorithm terminates after $L = L_{max} = 3$ unsuccessful, successive proposals (at $iter = 4$) and the overall cost in terms of forward calls (i.e., evaluations of $y(\Psi)$ and its derivative) was 200. Since forward model calls are required everytime any $\boldsymbol{\mu}_j$ is updated (Equation (4.34)), we plot the evolution of $\mathcal{F}$ (Equation (4.33)) with respect to the total number of $\boldsymbol{\mu}$-updates (including those for components that end up being deleted) in Figure 4.7.

To verify the results we carry out importance sampling as described in Section 2.2.3. The effective sample size (Equation (2.11)) was $ESS_{IS} = 0.96$, which is very close to 1. In Figure 4.8, the approximate posterior (Equation (4.38)) is compared with the exact posterior (IS), and excellent agreement is observed. One can see that not only the locations (mean) and the variances of the mixture components are captured correctly but also their corresponding probability weights.

For comparison purposes, and as the cost per forward model evaluation in this problem is negligible, we also performed random-walk MCMC with a Gaussian proposal density with standard deviation 0.35 that yielded an average acceptance ratio of 20%. The results are depicted in Figure 4.9. The corresponding ESS was $ESS_{MCMC} = 1 \times 10^{-3}$, i.e., roughly 1000 times more expensive (in terms of forward model evaluations) than the proposed strategy.

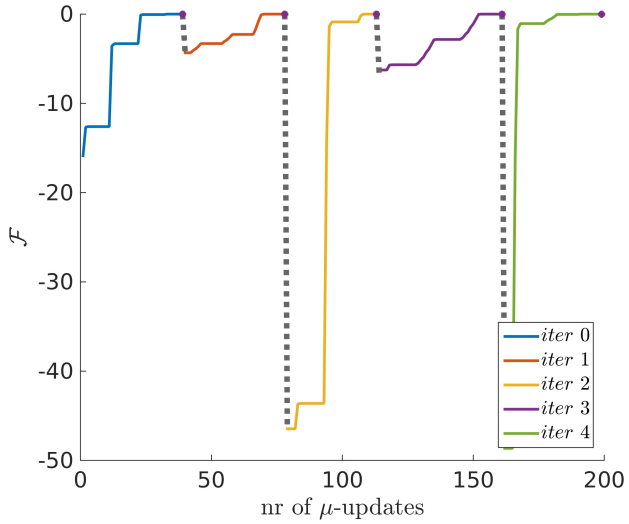


Figure 4.7: Evolution of $\mathcal{F}$ (Equation (4.33)) over the number of $\boldsymbol{\mu}$-updates (which is equal to the number of forward calls) for Example 1. Each color corresponds to a different value of $iter$. The number of $\boldsymbol{\mu}$-updates associated with mixture components that are subsequently deleted, is also included.

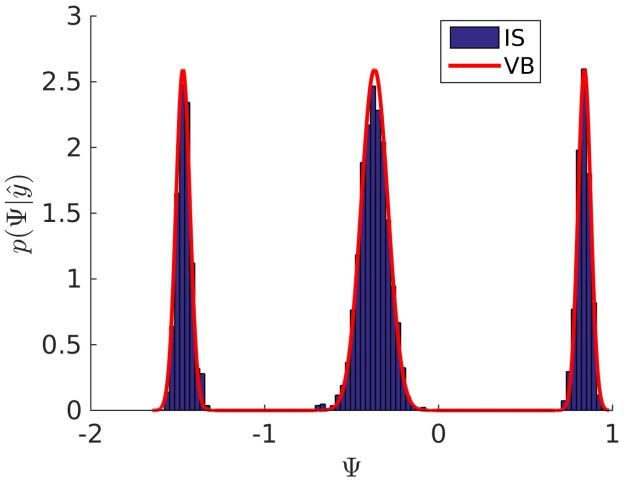
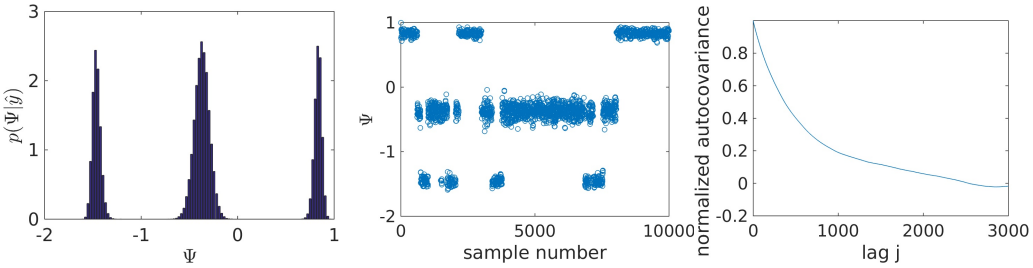


Figure 4.8: Exact (IS) and approximated (VB) posterior probability distribution, which show excellent agreement.

Elastography

In the motivating problem of nonlinear elastography, we simulate a scenario of applying a quasi-static pressure (e.g., with the ultrasound wand) and using the pre- and post-compression images to infer the material properties of the underlying tissue. We consider a two-dimensional domain $\Omega_0 = [0, 50] \times [0, 50]$, shown in Figure 4.10. The



(a) Normalized histogram. (b) Ψ over different samples (part of it). (c) Autocovariance.

Figure 4.9: (a): Posterior distribution obtained with random walk MCMC with 10^6 MCMC samples which coincides with Figure 4.8. (b): Evolution of the state Ψ per MCMC step. (c): Normalized autocovariance which decays slowly and results in a small ESS_{MCMC}.

governing equations consist of the conservation of linear momentum[7]:

$$\bigtriangledown \cdot (\boldsymbol{F}\boldsymbol{S}) = 0 \quad \text{in} \quad \Omega_0, \tag{4.56}$$

where $\boldsymbol{F} = \boldsymbol{I} + \nabla \boldsymbol{u}$ is the deformation map, $\boldsymbol{u}$ is the displacement field and $\boldsymbol{S}$ is the second Piola-Kirchhoff stress as described more in detail in Section 2.3. We assume Dirichlet boundary conditions along the bottom boundary (Figure 4.10), i.e.:

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } x_1 = [0, 50],\ x_2 = 0, \tag{4.57}$$

and the following Neumann conditions on the remaining boundaries:

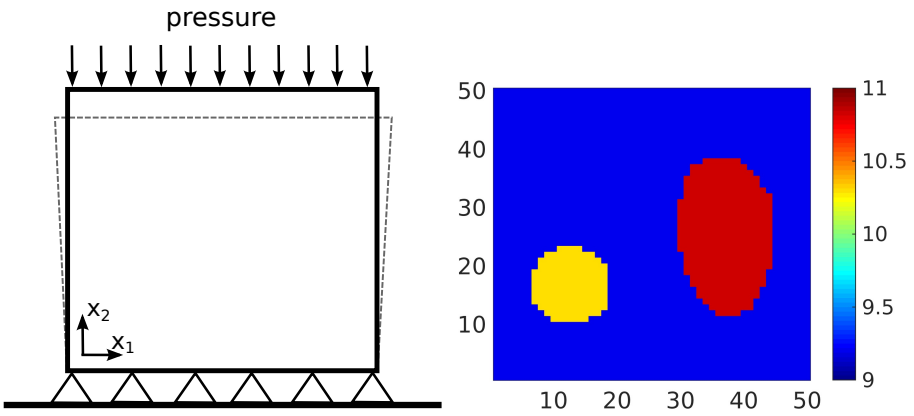
$$\begin{aligned} \boldsymbol{F}\boldsymbol{S} \cdot \boldsymbol{N} &= \begin{bmatrix} 0 \\ -100 \end{bmatrix}, \quad \text{on } x_1 \in [0, 50],\ x_2 = 50, \\ \boldsymbol{F}\boldsymbol{S} \cdot \boldsymbol{N} &= \boldsymbol{0}, \quad \text{on } x_1 = 0 \text{ and } x_1 = 50,\ x_2 \in [0, 50]. \end{aligned} \tag{4.58}$$

A nonlinear, elastic constitutive law (stress-strain relation) is adopted of the form:

$$\boldsymbol{S} = \frac{\partial w}{\partial \boldsymbol{E}}, \tag{4.59}$$

where $\boldsymbol{E} = \frac{1}{2}(\boldsymbol{F}^T\boldsymbol{F} - \boldsymbol{I})$ is the Lagrangian strain tensor and $w(\boldsymbol{E}, \psi)$ (Equation (2.29)) is the strain energy density function which depends (apart from $\boldsymbol{E}$) on the material parameters. In this example we employ the St. Venant-Kirchhoff model [140, 208, 209] that corresponds to the following strain energy density function w:

$$w = \frac{\nu\psi}{2(1+\nu)(1-2\nu)}[tr(\boldsymbol{E})]^2 + \frac{\psi}{2(1+\nu)}\, tr(\boldsymbol{E}^2). \tag{4.60}$$



(a) Problem configuration.

(b) Reference configuration of ψ in the log scale.

Figure 4.10: Problem and reference configuration.

The St. Venant-Kirchhoff model is an extension of the linear elastic material model to the nonlinear regime, i.e., large deformations. In this example $\nu = 0.3$ and the Young modulus ψ is assumed to vary in the problem domain, i.e., $\psi(\boldsymbol{x})$. In particular we assume the presence of two inclusions (tumors, Figure 4.10). In the larger, elliptic inclusion the Young modulus is $\psi = 50000$ (red), in the smaller, circular inclusion $\psi = 30000$ (yellow/orange) and in the remaining material $\psi = 10000$ (blue). The contrast $\frac{\psi_{inclusion}}{\psi_{matrix}} \approx 4-5$ coincides with experimental evidence on actual tissue [15, 16]. We generate synthetic data $\hat{\boldsymbol{y}}$ by using a 100×50 mesh and collecting the displacements at the interior points. These are in turn contaminated by zero mean, isotropic, Gaussian noise resulting in a signal-to-noise-ratio (SNR) of 1000. The forward solver used in the solution of the inverse problem consists of a regular grid with 50×50 quadrilateral finite elements. We assume that within each finite element, ψ is constant, resulting in a 2500 dimensional vector of inverse-problem unknowns $\boldsymbol{\Psi}$, $d_\Psi = 2500$. We note that in the discretized form, the resulting algebraic equations are nonlinear (geometric and material nonlinearities) and the state vector (forward-problem unknowns), i.e., the displacements, are of dimension 5100. Details about the implemented software can be found in Appendix E.

As in Chapter 3 for each mixture component we employ an adaptive learning scheme for the reduced coordinates Θ_i which are added one-by-one in such a way that they have a posteriori progressively smaller variances. For that reason we define the prior precisions $\lambda_{0,s,i}$ such that they are gradually larger. Given $\lambda_{0,s,1}$, which is assumed to be the same for all mixture components s, we define the *prior* precisions as follows

[7] Dependencies on the spatial variables $\boldsymbol{x} \in \Omega_0$ have been suppressed for simplicity.

(Equation (3.51)):

$$\lambda_{0,s,i} = \max(\lambda_{0,s,1}, \lambda_{s,i-1} - \lambda_{0,s,i-1}), \quad i = 2, 3, \ldots, d_\Theta. \tag{4.61}$$

$\lambda_{s,i-1}$ corresponds to the posterior precision for the previous reduced coordinate Θ_{i-1} of the same component s. This implies that, a priori, the next reduced coordinate will have at least the precision of the previous one as long as it is larger than the threshold $\lambda_{0,s,1}$. For the prior of $\boldsymbol{\eta}$ we use $\lambda_{0,\eta,s} = \max_i(\lambda_{0,s,i})$ as $\boldsymbol{\eta}$ represents the residual variance which is a priori smaller than the smallest variance of the reduced coordinates $\boldsymbol{\Theta}$. The results presented in the following were obtained for $\lambda_{0,s,1} = 1$ for all s and the material parameters are plotted in log scale.

The algorithm is initialized with four components, i.e., $S_0 = 4$, and $\Delta S = 3$ new components are proposed at each iteration $iter$ (Algorithm 3). Figure 4.11 depicts the mean $\boldsymbol{\mu}_1$ identified upon convergence ($iter = 0$) of an active component. Furthermore, it shows three perturbations, obtained according to Equation (4.43), which were used as initial values for the means of the $\Delta S = 3$ new components proposed at $iter = 1$.

Figure 4.12 depicts the evolution of the variational lower bound $\mathcal{F}$, (Equation (4.33)) per $\boldsymbol{\mu}$-update, i.e., per call to the forward model solver. In total the algorithm performed $iter = 24$ iterations which entailed proposing $S_0 + 24 \times \Delta S = 76$ new mixture components (until $L = L_{max} = 3$ was reached). For each of the 76 mixture components, the number of required forward calls ranged from 7 to 34. The total number of such calls was 1200.

Upon convergence, seven ($S = 7$) distinct mixture components were identified, which jointly approximate the posterior. The mean $\boldsymbol{\mu}_j$ of each component is shown in Figure 4.13 where the posterior responsibilities $q(s)$ are also reported. The numbering of the components relates to the order in which they were found by the algorithm. We observe that all mixture components identify the bulk of the two inclusions and most differences pertain to their boundaries (see also Figures 4.23, 4.24, 4.25). The shape of the boundaries has been found to play a defining role in distinguishing between malignant and benign tumors and metrics have been developed that provide a good diagnostic signature using this information [210],[211],[212]. Apart from the seven active components, the means of two additional mixture components ($s = 8, s = 9$) which were deactivated (based on the "Component Death" criterion in Algorithm 3), are shown.

In Table 4.3, we also report the (normalized) KL-divergence between all pairs of these nine components. One notes that component 8 was deleted because it was too similar to component 4 (from Equation (4.45) $d_{4,8} = 0.33 \times 10^{-2} < d_{min} = 0.01$) and component 9 was too similar to component 2 ($d_{2,9} = 0.56 \times 10^{-2} < d_{min} = 0.01$).

With regard to the covariance of each mixture component and the identification of the lower-dimensional subspaces, we employ the information-theoretic criterion previously discussed in order to adaptively determine the number of reduced-dimensions d_Θ.

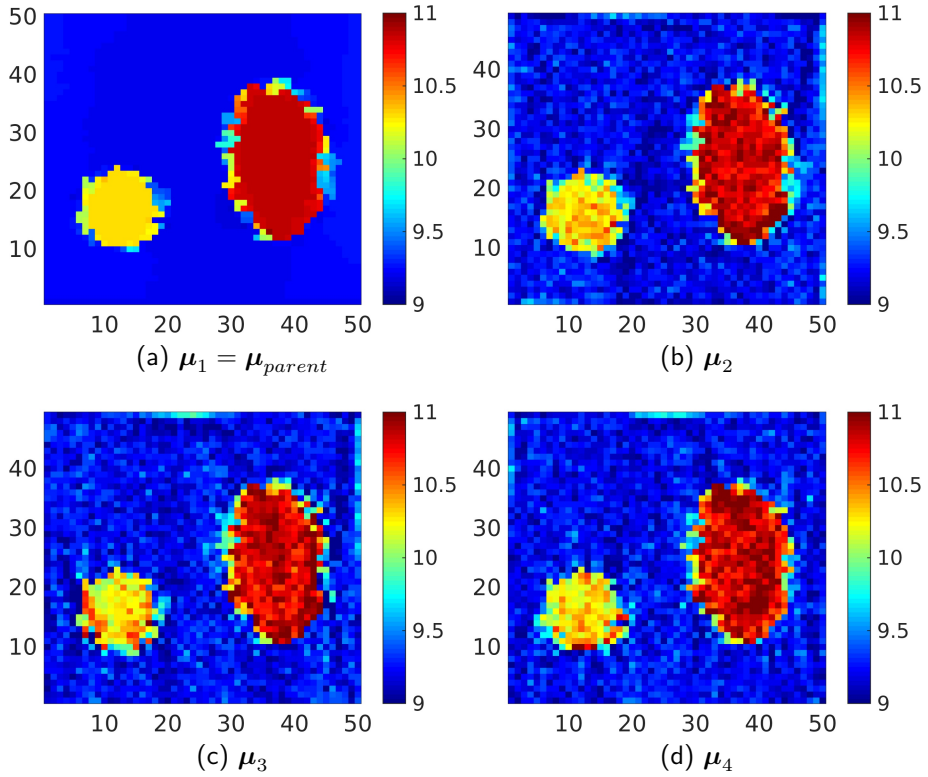


Figure 4.11: In (a) the converged $\boldsymbol{\mu}_1$ is depicted and in (b), (c) and (d) three perturbations (Equation (4.43)) used to initialize the means for the ΔS new proposed components (in log scale).

	s=1	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9
s=1	0	12.05	9.87	14.33	17.10	16.96	15.02	14.82	12.50
s=2		0	16.46	15.86	21.18	19.72	16.88	16.54	**0.56**
s=3			0	11.06	16.43	17.23	17.06	11.45	18.16
s=4				0	12.74	12.80	16.31	**0.33**	16.68
s=5					0	12.62	17.99	13.73	23.47
s=6						0	11.13	13.25	20.52
s=7							0	16.72	19.51
s=8								0	18.44

Table 4.3: Normalized KL divergences (Equation (4.45)) between all pairs of the mixture components. All values shown should be multiplied with $\times 10^{-2}$.

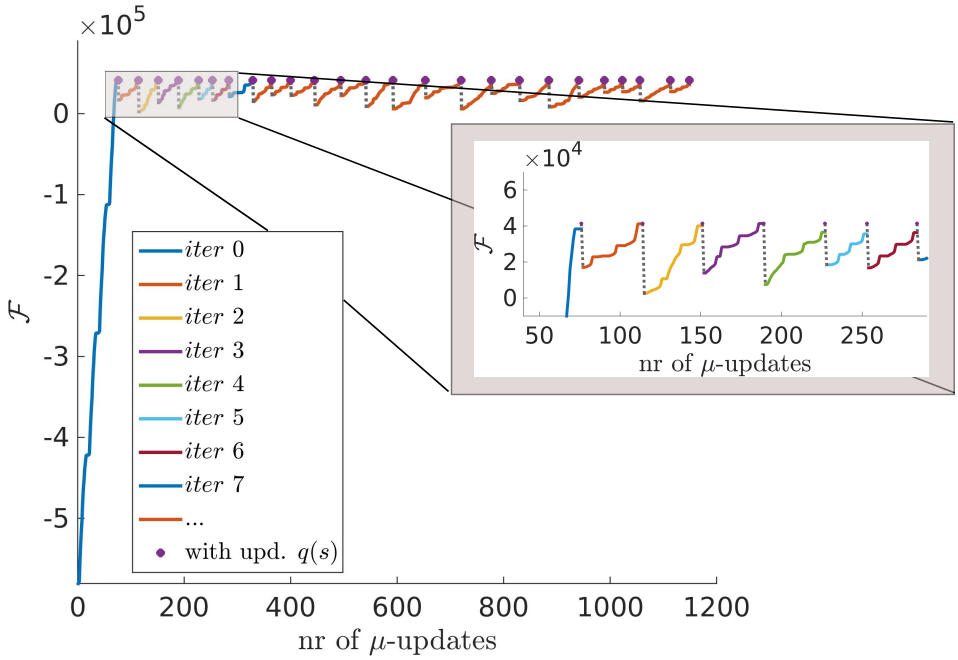


Figure 4.12: Evolution of $\mathcal{F}$ (Equation (4.33)) over the number of $\boldsymbol{\mu}$-updates (which is equal to the number of forward calls) for Example 2. Each color corresponds to a different value of $iter$. The number of $\boldsymbol{\mu}$-updates associated with mixture components that are subsequently deleted, is also included.

To that end, we use the relative information gains $I(d_\Theta, j)$ (Equation (4.36), see also Appendix C) which are depicted in Figure 4.14 for the three most active mixture components. We note that $I(d_\Theta, j)$ drops to relatively small values after a small number of reduced coordinates (with $d_\Theta = 8$, it drops to 1%). In the following results we used $d_\Theta = 11$. We discuss in Section 4.3 the behavior of the proposed scheme in cases in which the problem is not amenable to such a dimensionality reduction.

We defer further discussions on the individual mixture components in order to discuss the overall approximate posterior. The posterior mean and standard deviation of the mixture of Gaussians (Equation (4.40)) are shown in Figure 4.15. As expected, the posterior variance is largest at the boundaries of the inclusions.

Figure 4.16 depicts the posterior mean and $1\% - 99\%$ credible intervals along the diagonal of the problem domain, i.e., from $(0, 0)$ to $(50, 50)$. We note that the posterior quantiles envelop the ground truth.

For verification purposes we performed importance sampling as described in Section 2.2.3 in order to assess the overall accuracy of the approximation (a total of $M = 5000$ samples were generated). The effective sample size (Equation (2.11)) was $ESS_{IS} =$

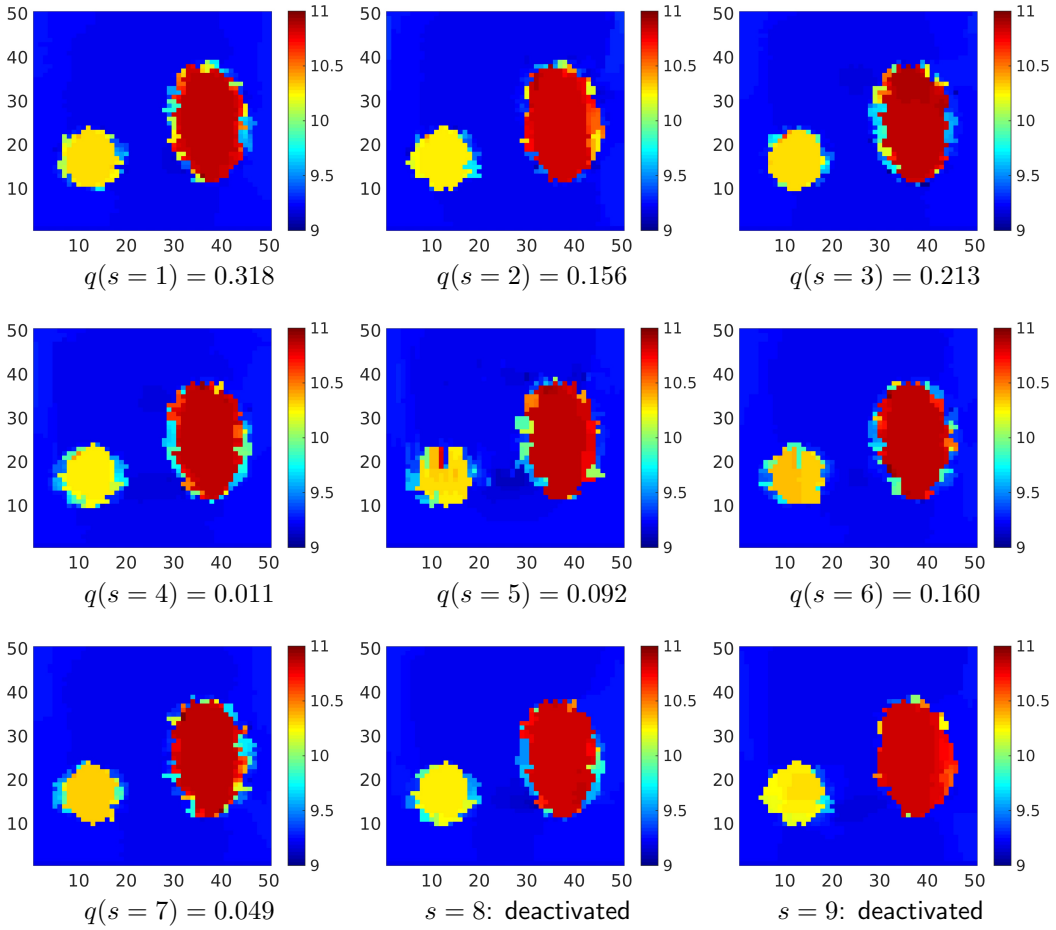


Figure 4.13: Posterior mean $\boldsymbol{\mu}_j$ for various mixture components in log scale and their posterior probabilities $q(s=j)$. The most active components are 1 and 3. Mixture components $8, 9$ are very similar to mixture components $4, 2$ respectively and are therefore deleted/deactivated (based on "Component Death" criterion in Algorithm 3, see also Table 4.3).

0.48 which indicates that the identified mixture of low-dimensional Gaussians provides a very good approximation to the actual posterior. In comparison, MCMC simulations performed using a Metropolis-adjusted Langevin scheme (MALA, [213]) exhibited very long correlation lengths resulting in $ESS_{MCMC} < 10^{-3}$.[8]

[8]Due to the computational expense, the MALA simulation results were actually obtained on a coarser discretization of the forward problem resulting in only 100 unknowns (in contrast to the 2500 in the target problem). The step sizes in the proposals were adapted to ensure that, on average, 60% of the moves were accepted [51]. The resulting ESS_{MCMC} was 10^{-3}. While additional fine- tuning could improve upon this, we doubt that for the actual problem which has 25 times more unknowns it

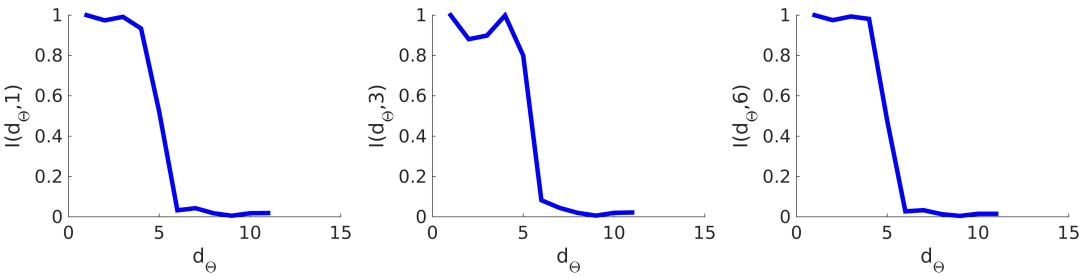


Figure 4.14: Information gain $I(d_\Theta, j)$ for three mixture components (see also Appendix C).

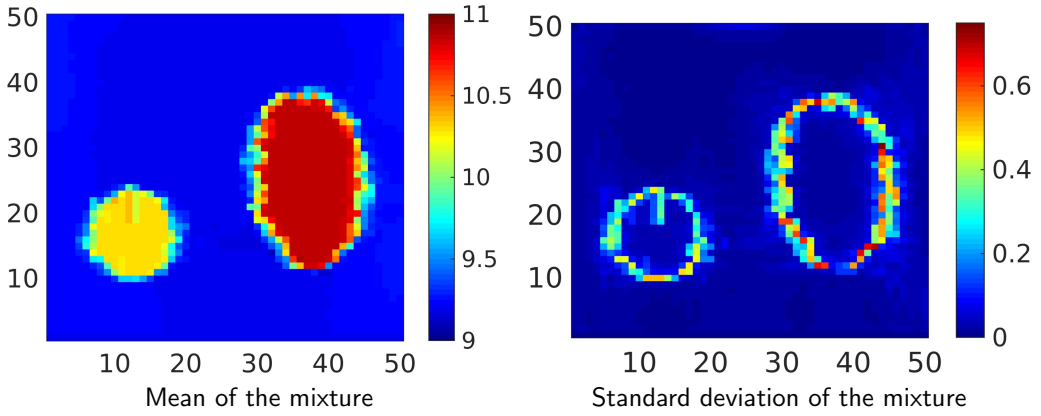


Figure 4.15: Approximate posterior mean and posterior standard deviation as computed from the mixture of Gaussians in Equation (4.40) (in log scale).

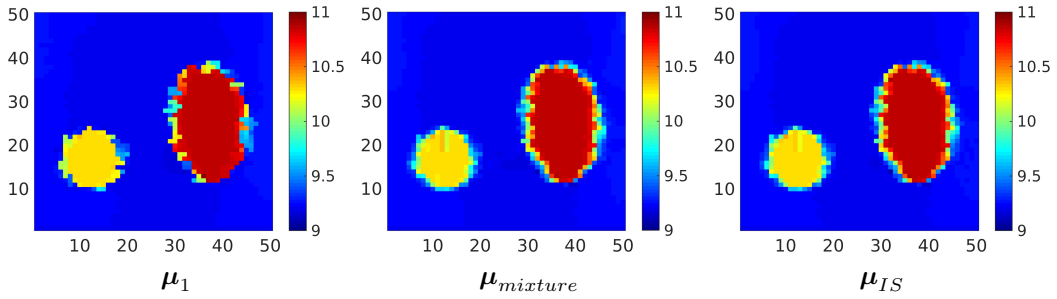


Figure 4.17: Comparison of the posterior mean found with a single mixture component (μ_1, left), with that found with a mixture of Gaussians ($\mu_{mixture}$, middle) and the exact mean estimated with IS (μ_{IS}, right). Depictions are in log scale.

will ever reach the ESS of the proposed approximation.

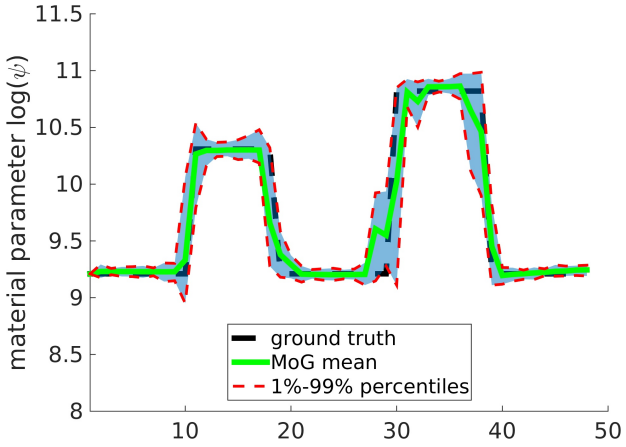


Figure 4.16: Posterior mean and credible intervals corresponding to 1% and 99%-(dashed lines), along the diagonal from $(0, 0)$ to $(50, 50)$.

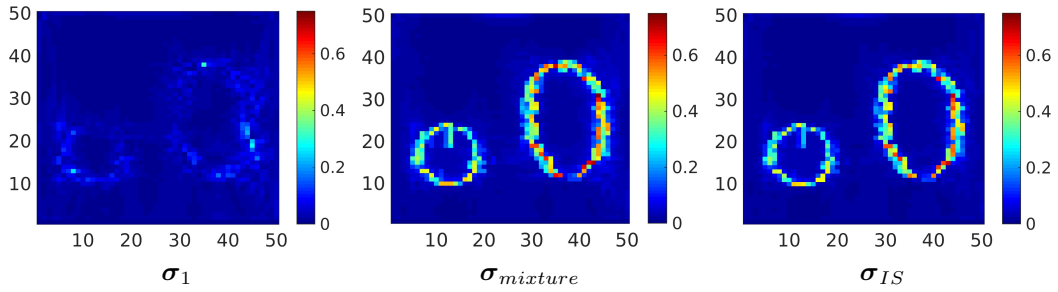


Figure 4.18: Comparison of the standard deviation of $\boldsymbol{\Psi}$ found with a single mixture component ($\boldsymbol{\sigma}_1$, left), with that found with a mixture of Gaussians ($\boldsymbol{\sigma}_{mixture}$, middle) and the exact values estimated with IS ($\boldsymbol{\sigma}_{IS}$, right). Depictions are in log scale.

In Figures 4.17 and 4.18, the approximate posterior mean and standard deviation are compared with the (asymptotically) exact values estimated by IS. Furthermore in these figures we plot the posterior mean and standard deviation found solely on the basis of the most prominent mixture component, i.e., $s = 1$. While, visually, the differences in the mean are not that striking (they are primarily concentrated at the boundaries of the inclusions), we observe that the posterior variance is clearly underestimated by a single component. In terms of the Euclidean norm (across the whole problem domain), we obtained that $\frac{||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{IS}||}{||\boldsymbol{\mu}_{mixture} - \boldsymbol{\mu}_{IS}||} = 5$ where $\boldsymbol{\mu}_{IS}$ is the exact mean obtained with IS and $\boldsymbol{\mu}_{mixture}$ is the approximate mean obtained from the mixture of Gaussians in Equation (4.40). Similarly for the standard deviation, we obtained that $\frac{||\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_{IS}||}{||\boldsymbol{\sigma}_{mixture} - \ bs\sigma_{IS}||} = 6$ where $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}, \boldsymbol{\sigma}_{IS}$ are the vectors of standard deviation across the whole problem domain, obtained with a single component, the mixture and IS respectively.

Figure 4.19 offers another view of the results along the diagonal of the problem domain and compares the 1% and 99% credible intervals with the (asymptotically) exact values found with IS.

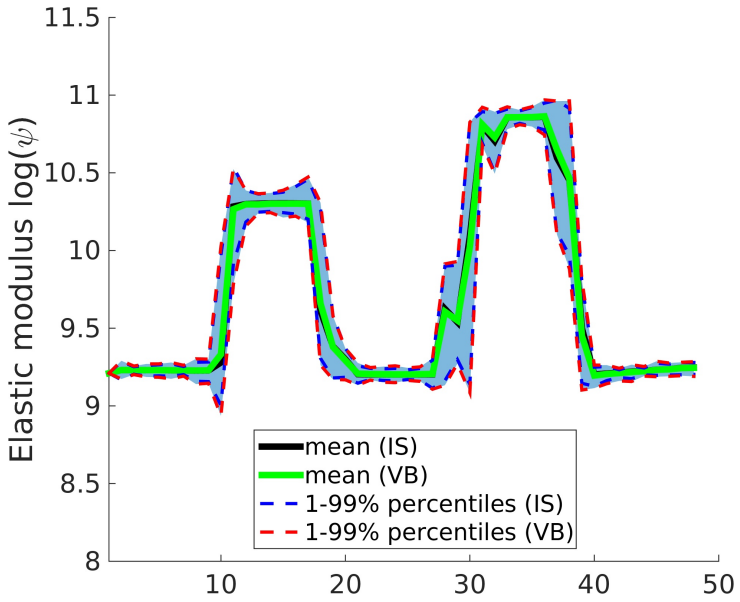


Figure 4.19: Posterior mean and credible intervals along the diagonal cut from $(0, 0)$ to $(50, 50)$ for mixture of Gaussians. Comparing the results with the results obtained by importance sampling (IS), we can see that they fit well to each other.

It is interesting to contrast this with Figure 4.20 which depicts the posterior along the same diagonal of the problem domain computed solely from each of the most prominent components, i.e., from $q_s(\boldsymbol{\Psi})$ for $s = 1, 3, 6$. We note again that away from the boundaries, strong similarities are observed, but none of the components by itself can fully capture or envelop the ground truth (compare also with Figure 4.16).

We provide further details on the most prominent mixture components, i.e., 1, 3 and 6. Figure 4.21 depicts the posterior standard deviation of $\boldsymbol{\Psi}$ as computed by using each of these components individually, i.e., from $q_s(\boldsymbol{\Psi})$ in Equation (4.38) for $s = 1, 3, 6$. All components yield small variance for the surrounding tissue and the interior of the inclusions while the posterior uncertainty is concentrated on the boundaries of the inclusions.

Figure 4.22 depicts the first four columns (basis vectors) of $\boldsymbol{W}_s$ for $s = 1, 3$ and the corresponding posterior variances $\lambda_{s,i}^{-1}$. The third column is perhaps the most informative, showing the differences between these vectors. These differences are most pronounced around the inclusions but, most importantly, reveal that the posterior variance is concentrated along different subspaces for different mixture components (see

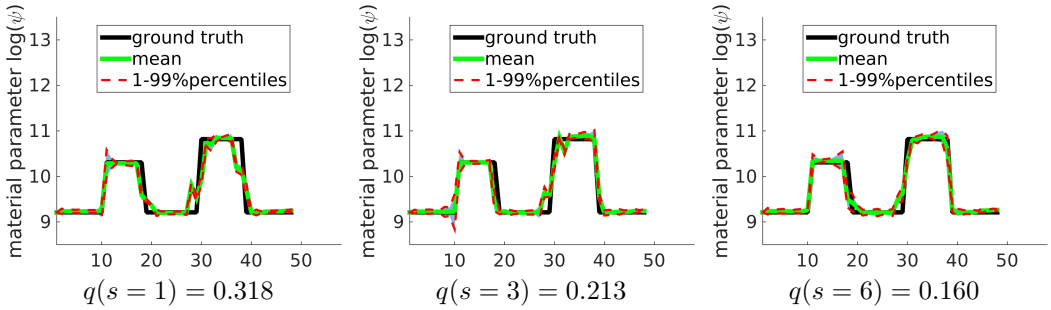


Figure 4.20: Posterior mean and $1\% - 99\%$ credibility intervals (dashed lines) along the diagonal cut from (0, 0) to (50, 50) for different mixture components.

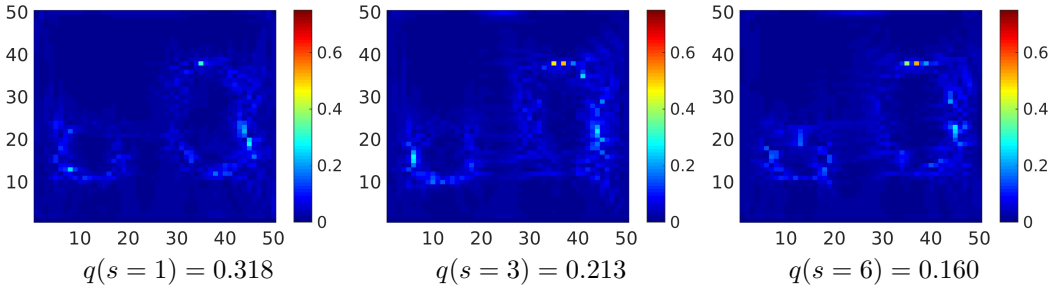


Figure 4.21: Standard deviation for selected mixture components in log scale.

also Figure 4.1). We note also with regard to $q(\boldsymbol{\eta}|s)$, i.e., the posterior of the residual noise in the representation of the unknowns, that for all mixture components, i.e., $s = 1, \dots 7$, $\lambda_{\eta,s}^{-1}$ was found approximately the same and equal to 4×10^{-3}, which is one or two orders of magnitude smaller than the variance associated with $\boldsymbol{\Theta}$ and has as a result a minimal overall influence.

In order to gain further insight we provide inference results along the boundary of the elliptical inclusion. In particular we consider the elements along the black line in Figure 4.23 and pay special attention to the elements marked with yellow stars from 1 to 4. We have purposely selected the black line to lie partly in the interior and partly in the exterior of the inclusion. In Figure 4.24, we plot the posterior mean along this black line (including credible intervals corresponding to 1% and 99%) as obtained exclusively from one of the three most prominent components (from $q_s(\boldsymbol{\Psi})$ in Equation (4.38) for $s = 1, 3, 6$) as well as by the mixture of Gaussians (Equation (4.37)). As it can now be seen more clearly, the individual components are only partially capable of capturing the ground truth. At times, points are misclassified in terms of whether they belong to the inclusion or not. On the other hand, the approximation provided by the mixture of Gaussians, averages over the individual components and leads to a

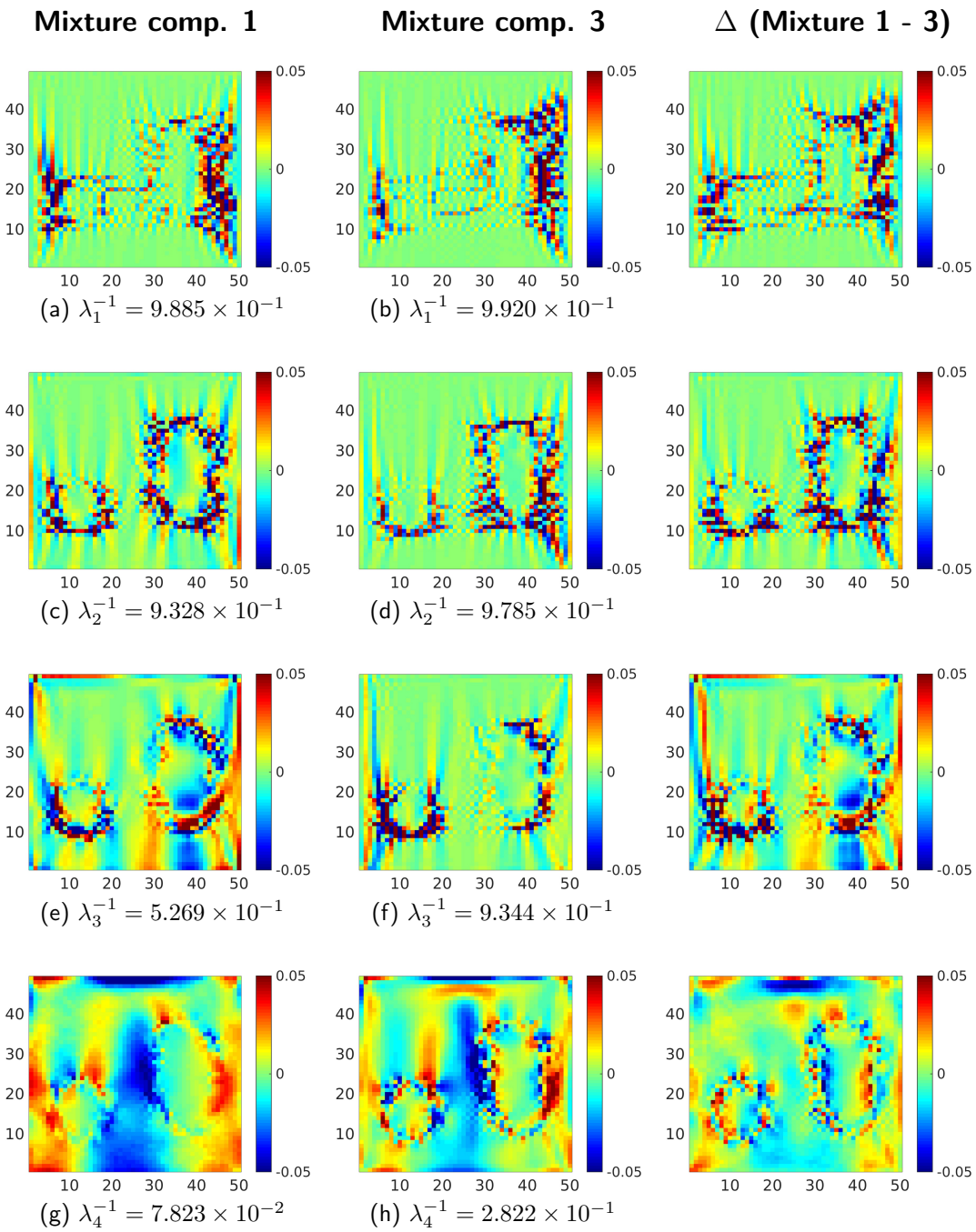


Figure 4.22: The first few basis vectors of $\boldsymbol{W}_j$ for mixture components $j = 1$ and $j = 3$ are shown in the first and second column. In the third column, the difference between the basis vectors in the first two columns, is plotted. The differences are more pronounced in the vicinity of the boundary of the inclusions.

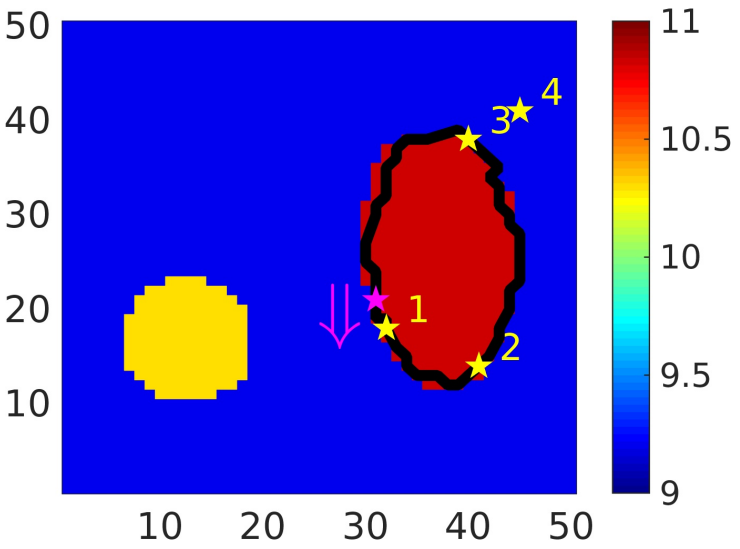


Figure 4.23: Posterior statistics for the elements along the black line are provided in Figure 4.24, starting at the magenta star and proceeding anti-clockwise around the inclusion. Posterior statistics for the elements marked by yellow stars $(1-4)$ are supplied in Figure 4.25. The background shows the ground truth in log scale.

posterior mean that is closer to the ground truth. More importantly, and especially in the regions where transitions from the inclusions to the surrounding tissue are present, the posterior uncertainty is larger to account for this ambiguity.

In Figure 4.25, we plot the posterior statistics for the elastic modulus of the elements 1 through 4 marked by yellow stars in Figures 4.23,4.24. The ground truth values are indicated with red rhombuses. We note that each of the mixture components gives rise to a Gaussian with, in general, a different mean/variance. These Gaussians reflect the uncertainty of the material properties at these points and are synthesized in the mixture. Interestingly, at element 4, which is further away from the boundary, all mixture components give rise to Gaussians with very similar means and variances, yielding a unimodal posterior when combined in the mixture. Three of the elements around the larger inclusion and their probabilistic behavior is shown in as well as one element of the surrounding tissue. One can see that some mixture components capture the exact values whereas others do not. However, the combination of mixture components includes the true value within its probability distribution.

Apart from the posterior probability distributions of the material parameters, we note also that noise precision was treated as an unknown and its (approximate) posterior was computed via Variational inference (Equation (4.24)). This is plotted in Figure 4.26.

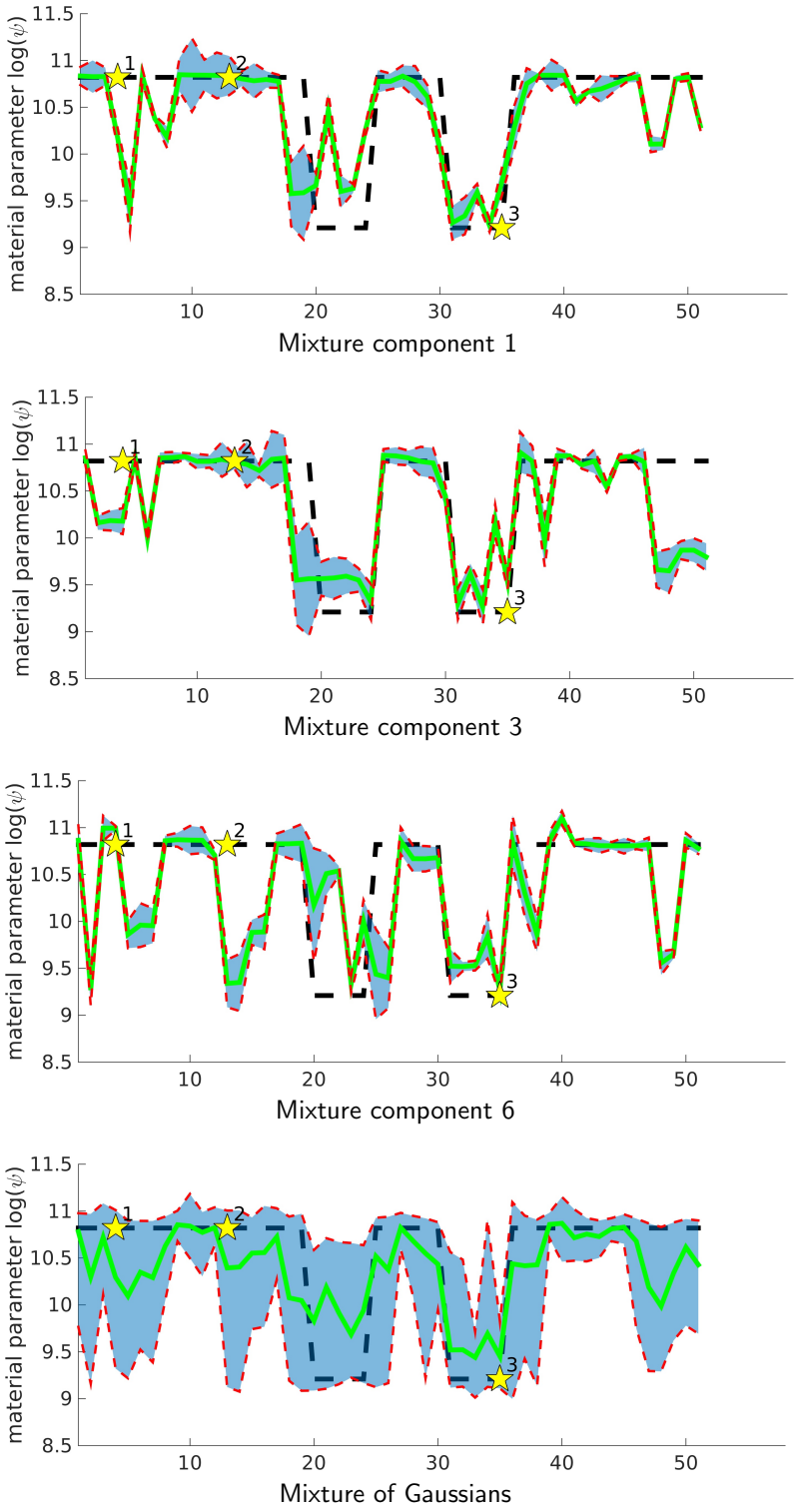


Figure 4.24: Posterior statistics along the black line of Figure 4.23. The ground truth is indicated by a black, dashed line. The posterior means are drawn with a green line — and credible intervals corresponding to 1%, 99% percentiles in red - - -.

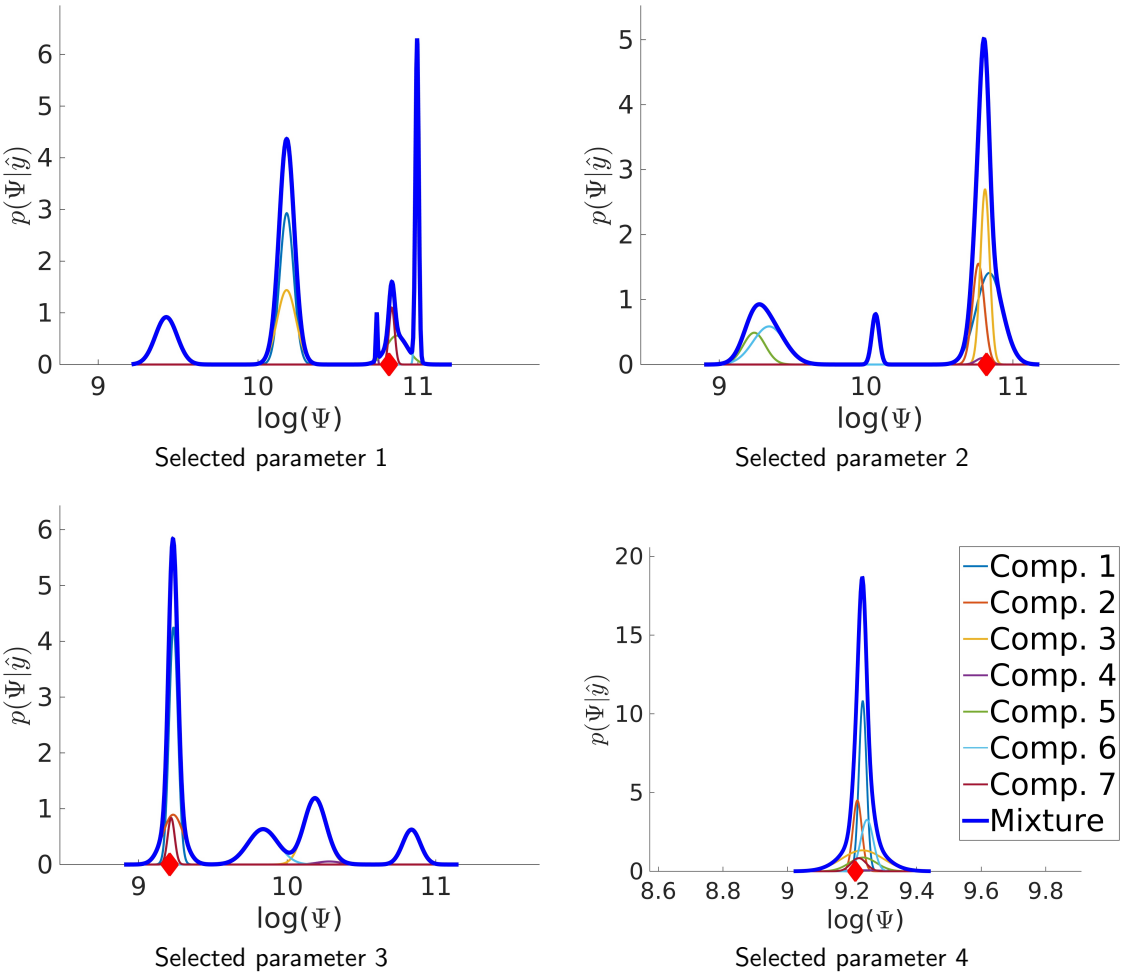


Figure 4.25: Posterior probability densities of the log of the elastic modulus, $\log(\Psi)$, of the elements 1 through 4 marked by yellow stars in Figure 4.23. The ground truth values are indicated with red rhombuses. The probability densities (Gaussians) associated with each of the 7 mixture components are multiplied by the corresponding posterior probabilities $q(s)$ and are shown by different colors. The combined, mixture of Gaussian is plotted in with a blue line.

Other examples/configurations

The previous results have demonstrated the capability of the proposed method not only to identify multiple modes but also to learn, for each mode, a different lower-dimensional subspace where posterior uncertainty is concentrated. There are of course problems where the posterior is either unimodal (or at least one mode is much more prominent than the rest) or such that the posterior variance is distributed equally over

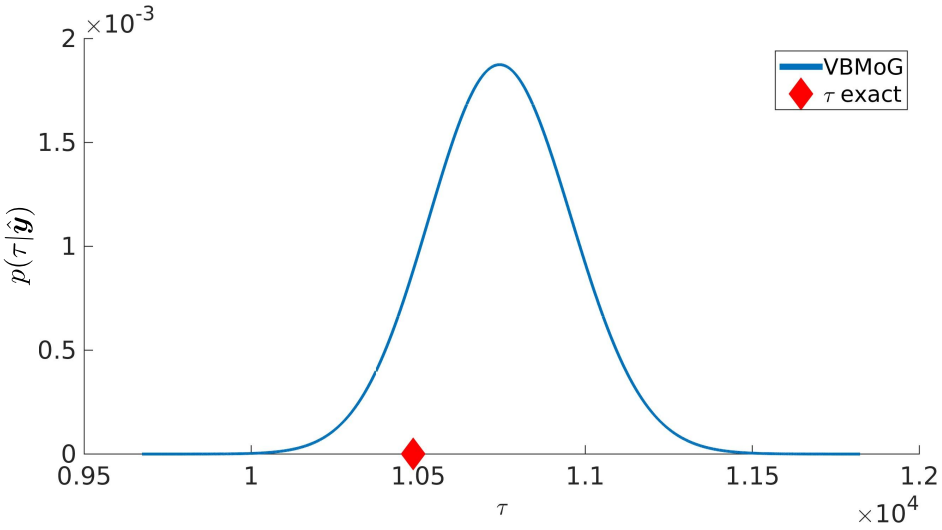


Figure 4.26: Approximate posterior $q(\tau)$ of the noise precision τ. The ground truth is indicated with the red rhombus.

a large number of dimensions (i.e., the posterior is not amenable to dimensionality reduction). In the context of the elastography problems examined the former scenario can take place when the noise in the data is relatively low. Then the data provide very strong evidence that precludes or make the presence of multiple modes of comparable significance unlikely.

The second scenario can appear when the available data is limited and/or very noisy. In this case, even if the posterior consists of a single mode, it is very likely that a large number of directions (if not all) will be characterized by large posterior variance as the signature of the data is very weak. In the following two subsections we examine such settings in order to demonstrate the ability of the proposed framework to adapt and provide good approximations even though these might consist of a single mode or they do not encompass any significant dimensionality reduction.

Example 2a: Only dimensionality reduction

We consider the same problem (i.e., the same material properties and forward model) but instead contaminate the data with much less noise resulting in a SNR of 1×10^4 (in contrast to 1×10^3 previously). In such a case a single mixture component is found. Despite multiple proposals (a total of 100 were attempted) the identified components are either deleted because they violate the KL-based similarity criterion (Equation (4.44)) or they have negligible posterior probability $q(s) \ll q_{min} = 10^{-3}$. The Gaussian identified has a mean that is extremely close to the ground truth as it

can be seen in Figure 4.27. The posterior variance across the problem domain is much smaller than in the previous setting (Figure 4.27) and is, as expected, due to the low levels of noise, concentrated along very few dimensions. Hence, the information gain metric $I(d_\Theta)$ decays extremely rapidly and we can adequately approximate the posterior with less than 5 reduced coordinates $\boldsymbol{\Theta}$ (Figure 4.27).

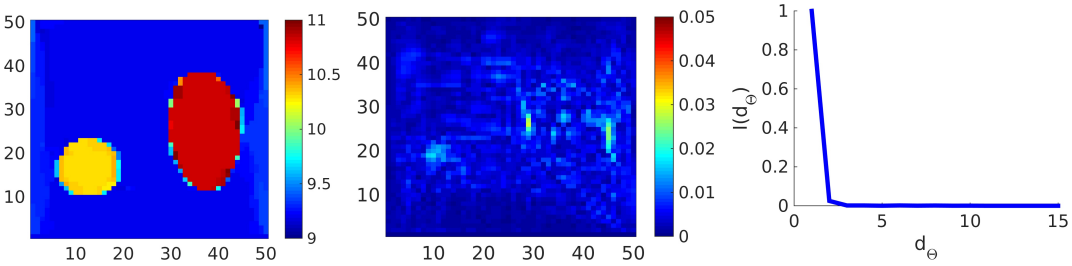


Figure 4.27: Example 2a: On the left panel, the posterior mean of the material parameters is plotted and in the middle panel the posterior standard deviation (in log scale). The right panel depicts the information gain as a function of d_Θ.

Example 2b: Only multimodality

We consider again the same problem (i.e., the same material properties and forward model) but instead contaminate the data with much more noise resulting in a SNR of 5×10^2 (in contrast to 1×10^3 previously) and assume that only half of the displacements are available, i.e., $d_y = 2550$ (in contrast to $d_y = 5100$ before). The proposed algorithm was employed and identified 21 active mixture components (in contrast to the 7 before). The means $\boldsymbol{\mu}_j$ of all these components are depicted in Figure 4.28) where the posterior probabilities $q(s)$ are also reported. As expected, the presence of more noise and the reduction in the available data have led to more modes in the posterior.

Moreover, as one would expect, none of these modes is particularly amenable to dimensionality reduction as the posterior variance is large and distributed along multiple dimensions. In fact by employing the information gain metric (Figure 4.29) we found that for most modes at least $d_\Theta \approx 750$ reduced coordinates were necessary to represent the variance accurately.

Nevertheless, the posterior mean estimated from the mixture of these 21 Gaussians (Equation (4.40)) is very close to the ground truth, see Figure 4.30. Understandably however, the posterior variance across the problem domain (Figure 4.30) is much larger.

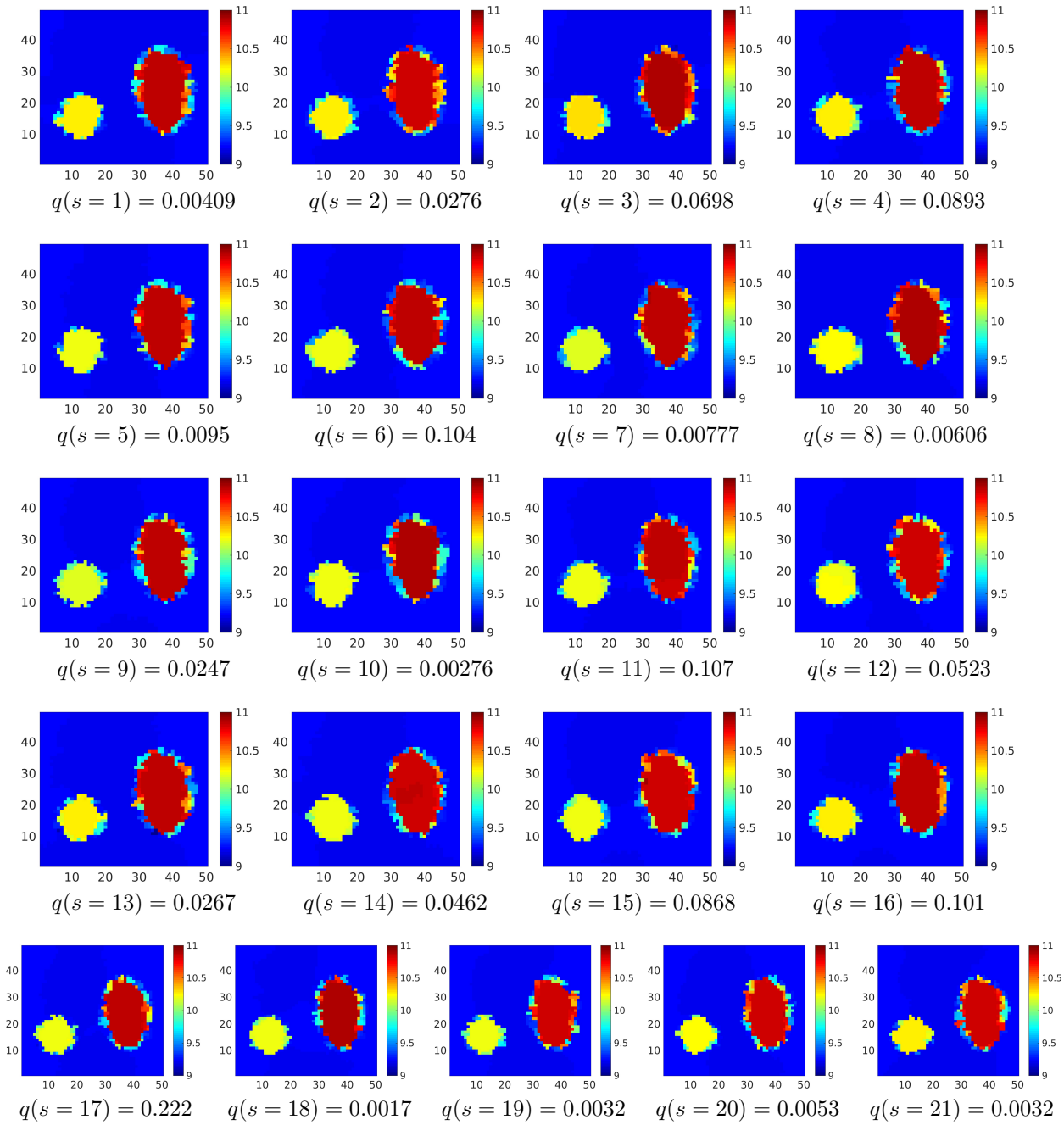
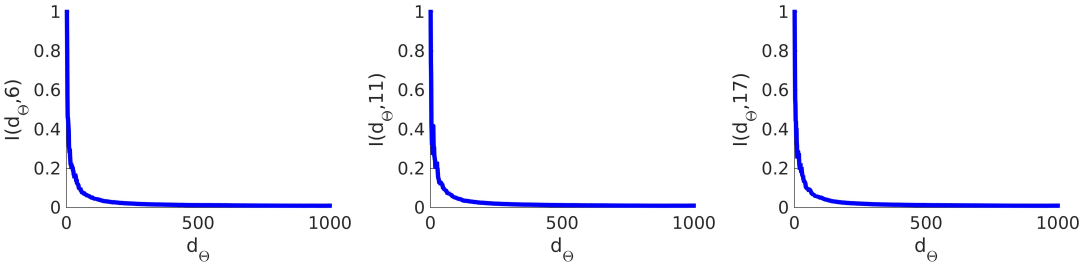


Figure 4.28: Example 2b: Posterior mean $\boldsymbol{\mu}_j$ and posterior probabilities $q(s = j)$ of each of the $S = 21$ mixture components identified (in log scale).



$q(s = 6) = 0.104$ $q(s = 11) = 0.107$ $q(s = 17) = 0.222$

Figure 4.29: Information gain $I(d_\Theta, j)$ for the 3 (out of the 21) mixture components with the largest posterior probability $q(s)$.

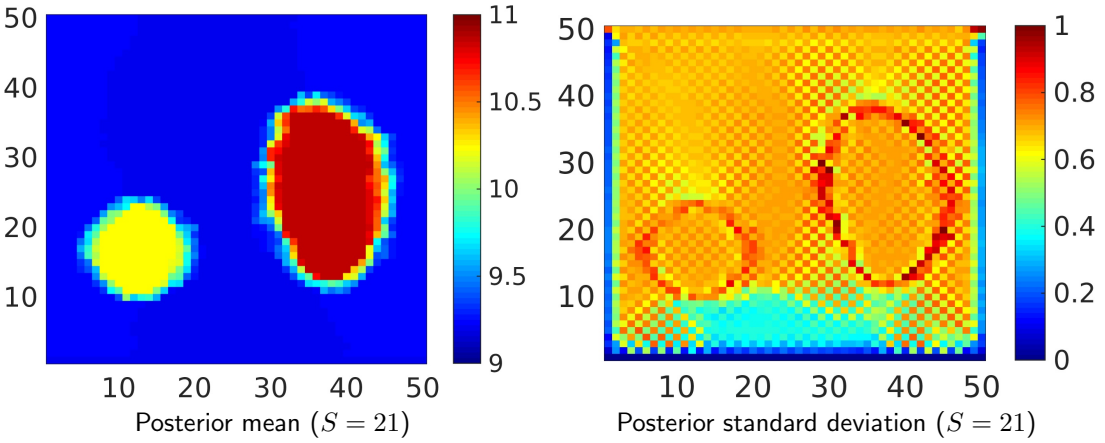


Figure 4.30: Example 2b: On the left panel, the posterior mean of the material parameters is plotted and in the right panel the posterior standard deviation (in log scale).

4.4 Summary

We presented a novel Variational Bayesian framework for the solution of inverse problems with computationally-demanding forward models and high-dimensional vector of unknowns. The strategy advocated addresses two fundamental challenges in the context of such problems. Firstly, the poor performance of existing inference tools in high-dimensions by identifying lower dimensional subspaces where most of the posterior variance is concentrated. Secondly, it is capable of capturing multimodal posteriors by making use of a mixture of multivariate Gaussians. Each of the Gaussians is associated with a different mean and covariance and provides an accurate local approximation. We verified the proposed strategy with importance sampling and as demonstrated in the numerical examples, the bias introduced by the approximations is small and can be very efficiently corrected, i.e., with very few forward model calls.

In the context of the motivating application, i.e., static, nonlinear elastography, it was shown that the computed multimodal approximation can provide a more accurate picture to the analyst or medical practitioner from which better diagnostic decisions can be drawn. In particular, the model proposed can better capture the spatial heterogeneity of material parameters which is a strong indicator of malignancy in tumors [210]. This is especially manifested in the boundaries of the inclusions (tumors) which can be better classified as well as in quantifying their effect in the results.

The method advocated is applicable to other problems characterized by high-dimensional vectors of unknowns, such as those involving spatially varying model parameters. It does not make use of any particular features of the forward model solver and requires the computation of first-order derivatives which can be handled with an adjoint formulation. While the number of forward model solutions, which is the primary metric of computational efficiency in our setting, increases with the number of identified posterior modes and depends on the stopping criteria employed, we have demonstrated that a few hundred forward calls are usually enough in the applications of interest. Furthermore, our algorithm is able to handle unimodal posteriors as well as densities which are not amenable to dimensionality reductions, e.g., due to large noise or sparse data.

We finally note that a restriction of the uncertainty quantification strategy proposed pertains to the forward model itself. Another source of uncertainty, which is largely unaccounted for, is *model uncertainty*. Namely, the parameters which are calibrated, are associated with a particular forward model (in our case a system of (discretized) PDEs), but one cannot be certain about the validity of the model employed. In general, there will be deviations between the physical reality where measurements are made, and the idealized mathematical/computational description. A critical extension therefore, particularly in the context of biomedical applications, is in the direction of identifying sources of model error and being able to quantify them in the final results, compare following chapter.

Chapter 5

Quantification of constitutive model error

> *Taking a model too seriously is really just another way of not taking it seriously at all.*
>
> Andrew Gelman, 1965-today.

5.1 The underestimated issue of model error

In the previous chapters, an efficient Bayesian framework for high-dimensional inverse problems to quantify parametric and observation errors was proposed and discussed. Another source of uncertainty, which is largely unaccounted for, is model uncertainty. In standard inverse problem formulations, the parameters which are calibrated are associated with a particular forward model, implicitly assuming that the model perfectly describes reality. However, in many cases there are discrepancies between the physical reality, where observations are made, and the idealized mathematical description.

As the true model is usually not known, there are several approaches that try to deal with this issue, such as the minimum description length (MDL) [137], the Akaike information criterion (AIC) [96], the Bayesian information criterion (BIC) [97] or the Bayes factor [95]. They are used for model selection, by favoring simple models that fit the data well. These criteria, however, do not quantify the model error but compare different models which each other, disregarding the possibility that the model might be wrong.

One of the first and widely used approaches to explicitly account for model error is based on Kennedy and O'Hagan [46] and has been extended by many other authors

[98, 99, 47]. In this framework, an explicit model error term $\delta(\boldsymbol{\Psi})$, e.g., described by a Gaussian process, is added to the model output:

$$\hat{\boldsymbol{y}} = \boldsymbol{y}(\boldsymbol{\Psi}) + \delta(\boldsymbol{\Psi}) + \boldsymbol{z}. \tag{5.1}$$

As in the previous chapters, $\boldsymbol{z}$ is the measurement error and $\boldsymbol{\Psi}$ denotes the unknown model parameters. This model is embedded in a larger framework, such as model calibration, i.e., parameter estimation, can be carried out. One advantage of the previous procedure is generality and that a deep understanding of the physical model is not required to identify its discrepancy. However, several drawbacks are noted. The discrepancy term $\delta(\boldsymbol{\Psi})$ is an empiricism, optimized with respect to the data. Therefore, it is tied to a particular quantity of interest and cannot be used for prediction. In addition, it is entangled with measurement errors and has a lack of physical insight [214].

To account for inadequate physics in the model, intrusive approaches have been developed, embedding an additive model error term within a submodel. One of the first approaches of this kind relates to Berliner [100], who improved one-dimensional ice-sheet models by adding an additional term to specific submodels. Alternatively, the calibration and validation problem can be reformulated as a kernel density estimation problem [174]. However, both approaches use a very small number of latent variables (less then 10 variables). Furthermore, the second approach employs a likelihood-free formulation and embeds the model error within the model parameters. The idea of embedding the model error in a submodel has been extended to high-dimensional problems in the field of fluid dynamics. To approximate turbulence models, such as RANS or the Spalart-Allmaras (SA) turbulence model, an additional term is added to account for the model error introduced by the model's approximative nature [104, 105, 106, 101, 102, 103]. The model error is computed by comparing the results of the approximate model with those of the 'true' but computationally very expensive model. The latter one obtains via direct numerical simulation (DNS). Within the discussed examples only the model error is treated as an unknown and no further latent or model parameters are quantified. This makes the problem a model calibration problem where the model error plays the role of a model parameter. Another drawback is that the results of the 'true' model have to be known, which is not the case in many applications.

In this thesis, we discuss a novel intrusive framework for model error estimation, which unfolds the classical forward problem to quantify model error in a physical manner. We propose to evaluate the model error within the constitutive model while fulfilling important physical laws, e.g., the conservation of linear momentum. This framework is based on a novel algorithm proposed by Koutsourelakis [108]. We extend this framework with a consistent derivation of the normalization term for a flexible integration of prior information. In addition, we propose a Variational Bayesian Expectation

Maximization algorithm to efficiently quantify the uncertainties of high-dimensional inverse problems. The presented methodology is based on a fully Bayesian formulation and validated by Gibbs sampling.

The remaining part of the chapter is organized as follows: In Section 5.2 the governing equations of solid mechanics are briefly reviewed. Furthermore, we present how these equations are incorporated in a fully Bayesian formulation. We employ an Expectation-Maximization scheme for efficiently estimating the model and latent variables. Finally, in Section 5.3, we present numerical illustrations in the context of elastography and validate the results.

5.2 Methods

Governing equations

In this section, we discuss an intrusive novel framework for model calibration and model validation. It focuses on quantifying the model discrepancies in the *constitutive* equation. The framework is inspired by deterministic approaches, such as the constitutive relation error (CRE) or the modified constitutive relation error (MCRE). CRE was originally developed for model validation in FEM simulations [215] and has been extended to inverse problems associated with elastostatic, elastodynamic, viscoelastic or viscoplastic materials [216, 217, 218, 219, 220]. For a detailed overview of CRE, MCRE and variations thereof, we refer to [221, 222].

Within our application of elastography, the forward model is based on continuum mechanics. In Section 2.3.2, we shortly reviewed the governing equations and showed how to derive numerical formulations. In the following framework, the individual equations will play an important role and will therefore derived in detail. The *governing equations* are described in the context of linear elastostatics:

- The *conservation of linear momentum*, which refers to Newton's second law of motion, is well-founded and generally accepted. Although discretized versions of the PDE introduce a discretization error, this is a well-studied problem and we refer the interested reader to [131, 144]. In this chapter, we disregard the discretization error and focus on the model error instead. The conservation law (Equation (2.26)) in the deformed configuration is

$$\bigtriangledown \cdot \tilde{\boldsymbol{\sigma}} + \rho_0 \boldsymbol{b} = \boldsymbol{0} \quad \text{in} \quad \Omega. \tag{5.2}$$

- The governing equations are supplemented by boundary constraints at the Dirichlet Γ_u and the Neumann Γ_S boundary (compare Equations 2.28 and 2.27):

$$\tilde{\boldsymbol{\sigma}} \cdot \boldsymbol{n} = \hat{\boldsymbol{t}} \quad \text{on} \quad \Gamma_S \quad \text{and} \tag{5.3}$$

$$\boldsymbol{u} = \boldsymbol{u}_b \quad \text{on} \quad \Gamma_u. \tag{5.4}$$

- The *constitutive law* describes the relation between stresses and strains. For a linear elastic material it is:

$$\tilde{\boldsymbol{\sigma}} = \boldsymbol{D} : \boldsymbol{e} \quad \text{in} \quad \Omega, \tag{5.5}$$

  where $\boldsymbol{D}$ is the elasticity tensor and $\boldsymbol{e}$ the Euler-Almansi strain.

For the numerical implementation the governing equations are weakly enforced (Section 2.3.3). For economy of notation, we employ triangular finite elements, which results in constant strains and stresses within each element.[1]

The weak form of the partial differential equation (Equation (5.2)) and the boundary conditions with weighting functions $\boldsymbol{v}$ is (Equation (2.30)):

$$\int_\Omega (\bigtriangledown \cdot \tilde{\boldsymbol{\sigma}} + \rho_0 \boldsymbol{b}) \cdot \boldsymbol{v} \; d\Omega + \int_{\Gamma_S} (\hat{\boldsymbol{t}} - \boldsymbol{\sigma} \cdot \boldsymbol{n}) \cdot \boldsymbol{v} \; d\Gamma = 0. \tag{5.6}$$

With the Gauss divergence theorem the conservation law is weakly informed by:

$$\int_\Omega \tilde{\boldsymbol{\sigma}} : (\bigtriangledown \boldsymbol{v})^T \; d\Omega = \int_{\Gamma_S} \hat{\boldsymbol{t}} \cdot \boldsymbol{v} \; d\Gamma + \int_\Omega \rho_0 \boldsymbol{b} \cdot \boldsymbol{v} \; d\Omega. \tag{5.7}$$

Discretized with d_{FE} number of elements this can be reformulated in:

$$\hat{\boldsymbol{B}}^T \boldsymbol{\sigma} = \boldsymbol{f}, \tag{5.8}$$

where

- $\hat{\boldsymbol{B}} \in \mathbb{R}^{d_f \times d_S}$ is the linear gradient operator with $\hat{\boldsymbol{B}}^T = \sum_{e=1}^{d_{FE}} \boldsymbol{L}_e^T \boldsymbol{B}_e^T V_e$. V_e is the volume of a single element e, $\boldsymbol{B}_e \in \mathbb{R}^{d_{S_e} \times d_f}$ the strain-displacement matrix for an element e and $\boldsymbol{L}_e \in \mathbb{R}^{d_{f_e} \times d_f}$ the Boolean matrix that relates local to global displacements and stresses. d_{S_e} represents the number of the stresses or strains per element.

- $\boldsymbol{\sigma} \in \mathbb{R}^{d_S}$ is the discretized stress tensor in vector form and d_S denotes its dimension. The entries of the stress vector relate to the constant stresses $\boldsymbol{\sigma}_e \in \mathbb{R}^{d_{S_e}}$ in an element e over d_{FE} elements: $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_1, ..., \boldsymbol{\sigma}_{d_{FE}}]^T$.

- $\boldsymbol{f} \in \mathbb{R}^{d_f}$ is the force vector with dimension d_f.

[1] If non-constant stress/strain elements are used, such as quadrilateral finite elements, a numerical integration within some of the equations is required.

The relation between stresses and strains for an element e is:

$$\boldsymbol{\sigma}_e = \boldsymbol{D}_e \boldsymbol{\epsilon}_e, \tag{5.9}$$

where

- the vector of strains $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_{d_{FE}}]^T \in \mathbb{R}^{d_S}$ contains the strains $\boldsymbol{\epsilon}_e$ of each elements e .

- $\boldsymbol{D}_e$ is the local constitutive matrix and

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{0} & ... & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2 & ... & \boldsymbol{0} \\ ... & ... & ... & ... \\ \boldsymbol{0} & \boldsymbol{0} & ... & \boldsymbol{D}_{d_{FE}} \end{bmatrix} \tag{5.10}$$

  the global constitutive matrix.

In the following, for the displacements $\boldsymbol{u} = (\boldsymbol{y} \cup \boldsymbol{u}_b) \in \mathbb{R}^{d_{y,all}}$ we distinguish unconstrained displacements $\boldsymbol{y} \in \mathbb{R}^{d_y}$ from displacements prescribed at the Dirichlet boundary $\boldsymbol{u}_b \in \mathbb{R}^{(d_{y,all}-d_y)}$. d_y is the number of unconstrained displacements and $d_{y,all}$ the total number of displacements.

The relation between strains and displacements is described by:

$$\boldsymbol{\epsilon} = \boldsymbol{B}\boldsymbol{u} = \boldsymbol{B}_y \boldsymbol{y} + \boldsymbol{B}_b \boldsymbol{u}_b, \tag{5.11}$$

where $\boldsymbol{B} \in \mathbb{R}^{d_S \times d_{y,all}}$ is the gradient operator, in this case the strain-displacement matrix. The entries in $\boldsymbol{B}_b$ refer to the Dirichlet boundary displacements and those in $\boldsymbol{B}_y$ to the unknown displacements:

$$\boldsymbol{B} = \sum_{e=1}^{d_{FE}} \boldsymbol{B}_e = \boldsymbol{B}_y \cup \boldsymbol{B}_b. \tag{5.12}$$

The strain-displacement relation on element level is:

$$\boldsymbol{\epsilon}_e = \boldsymbol{B}_{y,e} \boldsymbol{y}_e + \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e}, \tag{5.13}$$

where $\boldsymbol{y}_e$ and $\boldsymbol{u}_{b,e}$ are the unconstrained and constrained displacements for a specific element e.

Probabilistic formulation

In standard model calibration, the task is to quantify the unknown material parameters $\boldsymbol{\Psi}$, based on observations. In this chapter, we derive a novel Bayesian framework, which can address the two challenges from above at the same time: model calibration and validation. For this purpose, the governing equations are integrated in a Bayesian fashion. Therefore, we view displacements $\boldsymbol{y}$, stresses $\boldsymbol{\sigma}$, and material parameters $\boldsymbol{\Psi}$ as latent variables and incorporate them in the prior and likelihood.

We assume that we have measurements with white Gaussian noise available, leading to the **likelihood**:

$$p(\hat{\boldsymbol{y}}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}, \frac{1}{\tau}\boldsymbol{I}_{d_y}). \tag{5.14}$$

$\hat{\boldsymbol{y}}$ are the measured displacements and τ the measurement noise precision, which is assumed to be known.

To probabilistically incorporate the governing equations and further prior knowledge, we briefly summarize the required relations:

- The *conservation law* is modeled by a Gaussian with a very small variance, effectively enforcing the equilibrium of stresses.[2]

  $$\hat{\boldsymbol{B}}^T\boldsymbol{\sigma} - \boldsymbol{f} = \mathcal{N}(\boldsymbol{0}, \frac{1}{k}\boldsymbol{I}_{d_f}). \tag{5.15}$$

  k is the precision, treated as a constant, and will be raised iteratively during the simulations to enforce a stronger constraint.

- The *constitutive law* is modeled based on the following premise. When the constitutive law is not correct, the constitutive relation error $\boldsymbol{c}_e$ for each element e describes the discrepancy between the actual stresses, $\boldsymbol{\sigma}_e$, and the model-predicted stresses $\boldsymbol{D}_e\boldsymbol{\epsilon}_e$:

  $$\begin{aligned}\boldsymbol{\sigma}_e &= \boldsymbol{D}_e\boldsymbol{\epsilon}_e + \boldsymbol{c}_e \\ &= \boldsymbol{D}_e\boldsymbol{B}_e\boldsymbol{u}_e + \boldsymbol{c}_e \\ &= \boldsymbol{D}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}) + \boldsymbol{c}_e.\end{aligned} \tag{5.16}$$

  We assume that the model error can be modeled by a Gaussian distribution:

  $$\boldsymbol{c}_e = \mathcal{N}(\boldsymbol{0}, \frac{1}{v_e}\boldsymbol{I}_{d_{S_e}}). \tag{5.17}$$

  v_e is the unknown precision of the model error for element e. Large values for v_e correspond to elements with no/very small model error and small values for v_e to

[2] Alternatively, one could use an indicator function formulation to satisfy the conservation law exactly. Due to numerical issues, however, this has not been pursued.

elements with large model error. For each element e the precision of the model error can be different, following in $\boldsymbol{v} \in \mathbb{R}^{d_{FE} \times 1}$. Reformulating it with respect to $\boldsymbol{y}, \boldsymbol{\sigma}_e, \Psi_e$ and with Equation (5.19) it is:

$$p(\boldsymbol{y}, \boldsymbol{\sigma}_e, \Psi_e | v_e) \propto exp^{-\frac{v_e}{2}||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2}. \tag{5.18}$$

- The relation between inferred stresses and model-predicted stresses, described by the constitutive law, usually depends on material parameters. Assuming isotropic linear elasticity, the elastic modulus Ψ_e for each element e represents a linear scaling factor:

$$\boldsymbol{D}_e = \boldsymbol{D}_e(\Psi_e) = \Psi_e \tilde{\boldsymbol{D}}_e, \tag{5.19}$$

where $\tilde{\boldsymbol{D}}_e$ is known for a given Poisson ratio ν_e. The material parameters $\boldsymbol{\Psi} \in \mathbb{R}^{d_{FE}}$ are unknown and Ψ_e can be different for each finite element e, $e = 1 : d_{FE}$. The vector $\boldsymbol{\Psi}$ denotes the spatial discretization of the material parameters. We assume a smooth spatial distribution of $\boldsymbol{\Psi}$, i.e., we expect the constitutive properties to be locally correlated. For this purpose, we employ a hierarchical prior that penalizes jumps between neighboring elements:

$$p(\boldsymbol{\Psi}|\boldsymbol{H}) \propto -|\boldsymbol{H}|^{\frac{1}{2}} exp^{-0.5\boldsymbol{\Psi}^T(\boldsymbol{L}_\Psi^T \boldsymbol{H} \boldsymbol{L}_\Psi)\boldsymbol{\Psi}}, \tag{5.20}$$

where $\boldsymbol{L}_\Psi$ is a difference operator and $\boldsymbol{H} = diag(h_l)$ a diagonal matrix with dimension $d_L \times d_L$ and where d_L is the number of neighboring pairs.
We introduce an additional hyperprior for each parameter h_l, with:

$$p(h_l) = Gamma(a_{h,0}, b_{h,0}). \tag{5.21}$$

Then:

$$\begin{aligned} \log(p(\boldsymbol{\Psi}|\boldsymbol{H})\ p(\boldsymbol{H})) \quad &\propto \tfrac{1}{2}\textstyle\sum_{i=1}^{d_L} \log(h_i) - \tfrac{1}{2}\boldsymbol{\Psi}^T(\boldsymbol{L}_\Psi^T \boldsymbol{H} \boldsymbol{L}_\Psi)\boldsymbol{\Psi} \\ &\quad + \textstyle\sum_{i=1}^{d_L}[(a_{h,0} - 1)\log(h_i) - b_{h,0}h_i]. \end{aligned} \tag{5.22}$$

The prior knowledge of the constitutive and conservation law, as well as of the material parameters (compare Equations 5.17, 5.22, 5.15) is combined in a joint prior distribution:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v}) \quad &= \tfrac{\prod_{e=1}^{d_{FE}} p(\boldsymbol{y}, \boldsymbol{\sigma}_e, \Psi_e | v_e)\ p(\boldsymbol{\Psi}|\boldsymbol{H})\ p(\boldsymbol{H})\ p(\boldsymbol{\sigma})}{Z(\boldsymbol{v})} \\ &= \tfrac{\pi(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v})}{Z(\boldsymbol{v})}. \end{aligned} \tag{5.23}$$

$$Z(\boldsymbol{v}) = \int \pi(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v})\ d\boldsymbol{y}\ d\boldsymbol{\Psi}\ d\boldsymbol{H}\ d\boldsymbol{\sigma} \tag{5.24}$$

normalizes the unnormalized prior distribution

$$\pi(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v}) = \prod_{e=1}^{d_{FE}} p(\boldsymbol{y}, \boldsymbol{\sigma}_e, \Psi_e|v_e)\ p(\boldsymbol{\Psi}|\boldsymbol{H})\ p(\boldsymbol{H})\ p(\boldsymbol{\sigma}), \tag{5.25}$$

such that the resulting probability distribution $p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v})$ is a proper probability distribution, satisfying $\int p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v})\ d\boldsymbol{y}\ d\boldsymbol{\Psi}\ d\boldsymbol{H}\ d\boldsymbol{\sigma} = 1$. Of particular interest, within the normalization constant $Z(\boldsymbol{v})$, are the terms depending on $\boldsymbol{v}$, since those are required for updating $\boldsymbol{v}$. Lastly, we introduce a hyperprior $p_v(\boldsymbol{v})$ on the model error parameters $\boldsymbol{v}$.

Combining likelihood (Equation (5.14)) and prior (Equation (5.23)), the *posterior* density becomes:

$$p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}, \boldsymbol{v}|\hat{\boldsymbol{y}}) \propto p(\hat{\boldsymbol{y}}|\boldsymbol{y}) \frac{\pi(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v})}{Z(\boldsymbol{v})}\ p_v(\boldsymbol{v}). \tag{5.26}$$

Neglecting constant terms, the *log-posterior* follows:

$$\begin{aligned} \log(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}, \boldsymbol{v}|\hat{\boldsymbol{y}}) \approx\ & -\tfrac{\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2 \\ & - \textstyle\sum_{e=1}^{d_{FE}} \frac{v_e}{2}||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2 \\ & -\tfrac{k}{2}||\hat{\boldsymbol{B}}^T\boldsymbol{\sigma} - \boldsymbol{f}||^2 \\ & + \log\left(p(\boldsymbol{\Psi}|\boldsymbol{H})\ p(\boldsymbol{H})\right) \\ & - \log\ Z(\boldsymbol{v}) \\ & + \log\ p_v(\boldsymbol{v}). \end{aligned} \tag{5.27}$$

Bayesian inference

Traditional approaches derive a posterior of the form $p(\boldsymbol{\Psi}|\hat{\boldsymbol{y}})$. Compared to this, our posterior contains additional inference parameters $\boldsymbol{y}$, $\boldsymbol{\sigma}$, $\boldsymbol{H}$ and $\boldsymbol{v}$: $p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}, \boldsymbol{v}|\hat{\boldsymbol{y}})$. We also stress the fact that the classical forward model does not exist anymore. Instead, the new framework enables the quantification of the model inadequacy and the additional variables play the role of auxiliary variables. The *absence of a forward model* is particularly beneficial for solving high-dimensional problems, leading to reduced computational costs and higher efficiency. To explore the posterior and to make inferences about the unobserved parameters, Gibbs sampling [223] may be employed, or, for computational efficiency, Variational Bayes.

Since it is, in general, not possible to derive an explicit expression for $Z(\boldsymbol{v})$, an efficient and accurate VB scheme cannot be established. For this purpose, we employ, as discussed in the previous chapters, a hybrid, iterative scheme, based on Expectation-Maximization. While using the complete posterior distribution for inference on the parameters $\boldsymbol{\Upsilon} = \{\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}\}$, for $\boldsymbol{v}$ point estimates are derived. Maximum-a-Posteriori

(MAP) point estimates for $\boldsymbol{v}$, however, require the marginal posterior $p(\boldsymbol{v}|\hat{\boldsymbol{y}})$ given by:

$$p(\boldsymbol{v}|\hat{\boldsymbol{y}}) = \int p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}, \boldsymbol{v}|\hat{\boldsymbol{y}}) \; d\boldsymbol{y} \; d\boldsymbol{\Psi} \; d\boldsymbol{\sigma} \; d\boldsymbol{H}, \tag{5.28}$$

which is analytically intractable due to the coupling of the parameters $\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}$. The proposed iterative Expectation-Maximization scheme provides a remedy and is described in the following. Introducing the joint density $q(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}, \boldsymbol{v})$, a lower bound on the log-marginal posterior can be constructed:

$$\begin{aligned} \log p(\boldsymbol{v}|\hat{\boldsymbol{y}}) = \log \int p(\boldsymbol{v}, \boldsymbol{\Upsilon}|\hat{\boldsymbol{y}}) \; d\boldsymbol{\Upsilon} \geq \; & < \log \left(\frac{p(\boldsymbol{v}, \boldsymbol{\Upsilon}|\hat{\boldsymbol{y}})}{q(\boldsymbol{\Upsilon})}\right) >_{q(\boldsymbol{\Upsilon})} \\ = \; & < \log p(\hat{\boldsymbol{y}}|\boldsymbol{\Upsilon}) + \log \left(\frac{p(\boldsymbol{\Upsilon}|\boldsymbol{v})}{q(\boldsymbol{\Upsilon})}\right) >_{q(\boldsymbol{\Upsilon})} + \log \, p_v(\boldsymbol{v}) = \mathcal{F}(q(\boldsymbol{\Upsilon}), \boldsymbol{v}). \end{aligned} \tag{5.29}$$

Neglecting constant terms, this becomes:

$$\begin{aligned} \mathcal{F}(q(\boldsymbol{\Upsilon}), \boldsymbol{v}) \; = & < \log(p(\hat{\boldsymbol{y}}|\boldsymbol{y}) \; p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v}) \; p_v(\boldsymbol{v})) >_q - < \log \, q(\boldsymbol{\Upsilon}) >_q \\ & = -\tfrac{\tau}{2} < ||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2 >_q \\ & - \textstyle\sum_{e=1}^{d_{FE}} < \frac{v_e}{2} ||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e} \boldsymbol{y}_e + \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e})||^2 >_q \\ & - \tfrac{k}{2} < ||\hat{\boldsymbol{B}}^T \boldsymbol{\sigma} - \boldsymbol{f}||^2 >_q \\ & + < \log(p(\boldsymbol{\Psi}|\boldsymbol{H}) \; p(\boldsymbol{H})) >_q \\ & - \log \, Z(\boldsymbol{v}) + \log \, p_v(\boldsymbol{v}) - < \log \, q(\boldsymbol{\Upsilon}) >_q . \end{aligned} \tag{5.30}$$

The optimization of the lower bound with respect to its free parameters results in an iterative Expectation-Maximization scheme:

- **VB-Expectation**: Given $(\boldsymbol{v}^{(t-1)})$, find:

$$q^{(t)}(\boldsymbol{\Upsilon}) = \arg\max_q \mathcal{F}(q(\boldsymbol{\Upsilon}), \boldsymbol{v}^{(t-1)}), \tag{5.31}$$

- **VB-Maximization**: Given $q^{(t)}(\boldsymbol{\Upsilon})$, find:

$$\boldsymbol{v}^{(t)} = \arg\max_{\boldsymbol{v}} \mathcal{F}(q^{(t)}(\boldsymbol{\Upsilon}), \boldsymbol{v}), \tag{5.32}$$

representing a generalized coordinate ascent algorithm with respect to $\mathcal{F}$ (Figure 5.1). For $q(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H})$, we adopt a structured mean-field approximation (see Equation (2.19)) of the form:

$$q(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}) = q(\boldsymbol{y}) \; q(\boldsymbol{\Psi}) \; q(\boldsymbol{\sigma}) \; q(\boldsymbol{H}). \tag{5.33}$$

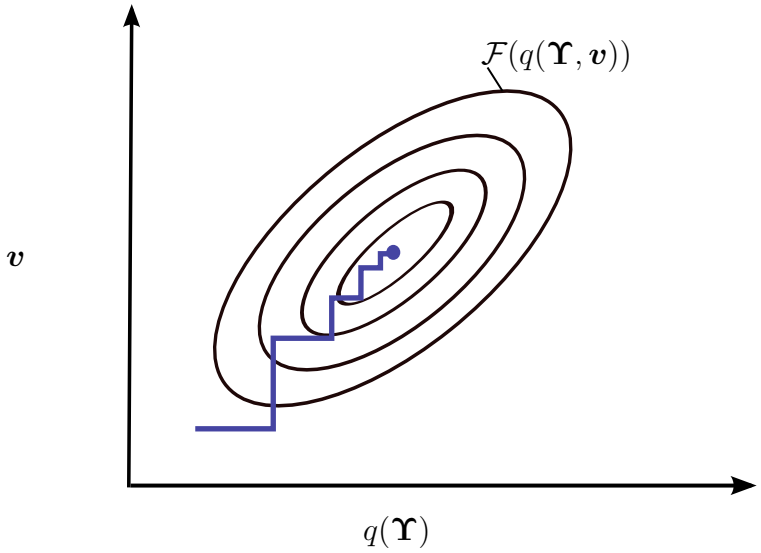


Figure 5.1: Schematic depiction of the Expectation-Maximization scheme for Variational Bayes.

5.2.1 Update equations for $q(\boldsymbol{y})$, $q(\boldsymbol{\Psi})$, $q(\boldsymbol{\sigma})$, $q(\boldsymbol{H})$ and $\boldsymbol{v}$

Based on the likelihood, priors and the adopted mean-field assumption, we can infer that the optimal approximate posteriors will be:

$$\begin{aligned} &q^{opt}(\boldsymbol{y}) \equiv \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Lambda}_y^{-1}), \\ &q^{opt}(\boldsymbol{\sigma}) \equiv \mathcal{N}(\bar{\boldsymbol{\mu}}_\sigma, \bar{\boldsymbol{\Lambda}}_\sigma^{-1}), \\ &q^{opt}(\boldsymbol{\Psi}) \equiv \mathcal{N}(\bar{\boldsymbol{\mu}}_\Psi, \bar{\boldsymbol{\Lambda}}_\Psi^{-1}), \\ &q^{opt}(\boldsymbol{H}) \equiv \prod_{l=1}^{d_L} Gamma(a_{h,l}, b_{h,l}). \end{aligned} \tag{5.34}$$

E-step - Variational Bayes

The update equation for the E-step - Variational Bayes can be readily established:

- For $q(\boldsymbol{\Psi}) = \mathcal{N}(\bar{\boldsymbol{\mu}}_\Psi, \bar{\boldsymbol{\Lambda}}_\Psi^{-1})$

$$\begin{aligned} \Lambda_{\Psi,e} &= v_e < (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})^T \tilde{\boldsymbol{D}}_e^T \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}) >_q \\ &= v_e(\tilde{\boldsymbol{D}}_e^T \tilde{\boldsymbol{D}}_e : ((\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})(\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})^T) \\ &+ v_e(\tilde{\boldsymbol{D}}_e^T \tilde{\boldsymbol{D}}_e : (\boldsymbol{B}_{y,e}\boldsymbol{\Sigma}_{y,e}\boldsymbol{B}_{y,e}^T)), \end{aligned} \tag{5.35}$$

$$\begin{aligned} \Lambda_{\Psi,e}\mu_{\Psi,e} &= v_e < (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})^T \tilde{\boldsymbol{D}}_e^T \boldsymbol{\sigma}_e) >_q \\ &= v_e(\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})^T \tilde{\boldsymbol{D}}_e^T \bar{\boldsymbol{\mu}}_{\sigma,e}. \end{aligned} \tag{5.36}$$

Including the smoothing prior (compare Equation (5.20)), the updated variance and mean are:

$$\bar{\boldsymbol{\Lambda}}_{\Psi} = \boldsymbol{\Lambda}_{\Psi} + \boldsymbol{L}_{\Psi}^{T} < \boldsymbol{H} >_{q} \boldsymbol{L}_{\Psi}, \tag{5.37}$$

and

$$\bar{\boldsymbol{\mu}}_{\Psi} = \bar{\boldsymbol{\Lambda}}_{\Psi}^{-1}(\boldsymbol{\Lambda}_{\Psi}\boldsymbol{\mu}_{\Psi}). \tag{5.38}$$

- For $q(\boldsymbol{H}) = \prod_{l=1}^{d_L} Gamma(a_{h,l}, b_{h,l})$
  The hyperparameters h_l, conditional on the other parameters, follow a Gamma distribution $Gamma(a_{h,l}, b_{h,l})$ with

$$a_{h,l} = a_{h,0} + \frac{1}{2}, \tag{5.39}$$

$$\begin{aligned} b_{h,l} &= b_{h,0} + \tfrac{1}{2} < (\Psi_{l,1} - \Psi_{l,2})^2 >_q \\ &= b_{h,0} + \tfrac{1}{2}(< \Psi_{l,1}^2 >_q - 2 < \Psi_{l,1}\Psi_{l,2} >_q + < \Psi_{l,2}^2 >_q) \\ &= b_{h,0} + \tfrac{1}{2}(\bar{\mu}_{\Psi_{l,1}}^2 - 2\bar{\mu}_{\Psi_{l,1}}\bar{\mu}_{\Psi_{l,2}} + \bar{\mu}_{\Psi_{l,2}}^2) \\ &+ \tfrac{1}{2}(\bar{\Sigma}_{\Psi,l,11} - 2\bar{\Sigma}_{\Psi,l,12} + \bar{\Sigma}_{\Psi,l,22}). \end{aligned} \tag{5.40}$$

- For $q(\boldsymbol{\sigma}) = \mathcal{N}(\bar{\boldsymbol{\mu}}_{\sigma}, \bar{\boldsymbol{\Lambda}}_{\sigma}^{-1})$
  On an element level, the constitutive relation is expressed by:

$$\boldsymbol{\Lambda}_{\sigma,e} = v_e \boldsymbol{I}_{d_{S_e}}, \tag{5.41}$$

$$\boldsymbol{\Lambda}_{\sigma,e}\boldsymbol{\mu}_{\sigma,e} = v_e < \Psi_e \tilde{\boldsymbol{D}}_e \boldsymbol{\epsilon}_e >= v_e \bar{\mu}_{\Psi,e} \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}). \tag{5.42}$$

Including the conservation law constraint equation, we obtain:

$$\bar{\boldsymbol{\Lambda}}_{\sigma} = k\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^{T} + \boldsymbol{\Lambda}_{\sigma}, \tag{5.43}$$

$$\bar{\boldsymbol{\mu}}_{\sigma} = \bar{\boldsymbol{\Sigma}}_{\sigma}(k\hat{\boldsymbol{B}}\boldsymbol{f} + \boldsymbol{\Lambda}_{\sigma}\boldsymbol{\mu}_{\sigma}). \tag{5.44}$$

- For $q(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Lambda}_y^{-1})$

$$\begin{aligned} \boldsymbol{\Lambda}_y &= \textstyle\sum_{e=1}^{d_{FE}} v_e < \boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}_e \Psi_e \Psi_e \tilde{\boldsymbol{D}}_e \boldsymbol{B}_{y,e} \boldsymbol{L}_{y,e} >_q + \tau \boldsymbol{I}_{d_y} \\ &= \textstyle\sum_{e=1}^{d_{FE}} v_e (\boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}_e \tilde{\boldsymbol{D}}_e \boldsymbol{B}_{y,e} \boldsymbol{L}_{y,e} (\bar{\mu}_{\Psi,e}^2 + \bar{\Sigma}_{\Psi,e}) + \tau \boldsymbol{I}_{d_y}, \end{aligned} \tag{5.45}$$

$$\boldsymbol{\mu}_y = \boldsymbol{\Lambda}_y^{-1}[\textstyle\sum_{e=1}^{d_{FE}} v_e < \boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}_e \Psi_e (\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e}) >_q + \tau \hat{\boldsymbol{y}}]. \tag{5.46}$$

Incorporating q^{opt} into the previous lower bound expression (Equation (5.30)) and neglecting constant terms, the lower bound $\mathcal{F}$ takes the following form:

$$\begin{aligned}\mathcal{F}(q^{opt}(\boldsymbol{\Upsilon}), \boldsymbol{v}) &= < \log(p(\hat{\boldsymbol{y}}|\boldsymbol{y})\ p(\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v})\ p(\boldsymbol{v})) >_q \\ &- < \log(q^{opt}(\boldsymbol{y})\ q^{opt}(\boldsymbol{\Psi})\ q^{opt}(\boldsymbol{H})\ q^{opt}(\boldsymbol{\sigma})) >_q \\ &= -\tfrac{\tau}{2}||\hat{\boldsymbol{y}} - \boldsymbol{\mu}_y||^2 - \tfrac{\tau}{2} tr(\boldsymbol{\Sigma}_y) \\ &- \textstyle\sum_{e=1}^{d_{FE}} < \frac{v_e}{2} ||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2 >_q - \log Z(\boldsymbol{v}) \\ &- \tfrac{k}{2}(||\hat{\boldsymbol{B}}^T \bar{\boldsymbol{\mu}}_\sigma - \boldsymbol{f}||^2 + tr(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T \bar{\boldsymbol{\Sigma}}_\sigma)) \\ &- \textstyle\sum_{j=1}^{d_L} a_{h,j} \log\ b_{h,j} + d_L \log(\Gamma(a_{h,j})) + \log p_v(\boldsymbol{v}) \\ &- \tfrac{1}{2}\log|\boldsymbol{\Lambda}_y| - \tfrac{1}{2}\log|\boldsymbol{\Lambda}_\Psi| - \textstyle\sum_{j=1}^{d_L} a_{h,j} \log\ b_{h,j} - \tfrac{1}{2}\log|\boldsymbol{\Lambda}_\sigma|.\end{aligned} \tag{5.47}$$

M-step

Within the VB-Maximization step, we want to derive point estimates for $\boldsymbol{v}$. Therefore, we examine the terms in $\mathcal{F}$ that depend on $\boldsymbol{v}$:

$$\begin{aligned}\mathcal{F}_v &= < \log\ \pi(\boldsymbol{u}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}|\boldsymbol{v}) >_q + \log\ p_v(\boldsymbol{v}) - \log Z(\boldsymbol{v}) \\ &\propto - \textstyle\sum_{e=1}^{d_{FE}} \frac{v_e}{2} < ||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2 >_q \\ &+ \log\ p_v(\boldsymbol{v}) - \log Z(\boldsymbol{v}).\end{aligned} \tag{5.48}$$

With the derivative

$$\begin{aligned}\tfrac{\partial \mathcal{F}_v}{\partial v_e} &= -\tfrac{1}{2} < ||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e \boldsymbol{B}_e \boldsymbol{L}_e \boldsymbol{u}||^2 >_q + \tfrac{\partial \log\ p_v(\boldsymbol{v})}{\partial v_e} - \tfrac{\partial \log\ Z(\boldsymbol{v})}{\partial v_e} \\ &= -\tfrac{1}{2}\Xi_e + \tfrac{\partial \log\ p_v(\boldsymbol{v})}{\partial v_e} - \tfrac{\partial \log\ Z(\boldsymbol{v})}{\partial v_e},\end{aligned} \tag{5.49}$$

the model error v_e can be obtained via gradient ascent:

$$v_e = v_e + \alpha \frac{\partial \mathcal{F}_v}{\partial v_e}. \tag{5.50}$$

The expectation Ξ_e can be computed from the E-step:

$$\begin{aligned}\Xi_e &= < ||\boldsymbol{\sigma}_e - \boldsymbol{D}_e \boldsymbol{\epsilon}_e||^2 >_q \\ &= < ||(\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}))||^2 >_q \\ &= (\bar{\boldsymbol{\mu}}_{\sigma,e}^T \bar{\boldsymbol{\mu}}_{\sigma,e} + tr(\bar{\boldsymbol{\Sigma}}_{\sigma,e})) - 2\bar{\boldsymbol{\mu}}_{\sigma,e}^T \mu_{\Psi,e} \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}) \\ &+ (\bar{\mu}_{\Psi,e}\bar{\mu}_{\Psi,e} + \bar{\Sigma}_{\Psi,e}) \tilde{\boldsymbol{D}}_e^T \tilde{\boldsymbol{D}}_e : [(\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})(\boldsymbol{B}_{y,e}\boldsymbol{\mu}_{y,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})^T \\ &+ (\boldsymbol{B}_e \boldsymbol{\Sigma}_{u,e} \boldsymbol{B}_e^T)].\end{aligned} \tag{5.51}$$

The partition function $Z(\boldsymbol{v}) = \int \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})\ d\boldsymbol{\Upsilon}$ (Equation (5.24)) which normalizes $\pi(\boldsymbol{\Upsilon}|\boldsymbol{v})$ is not readily deducible. However, one can reformulate with $p(\boldsymbol{\Upsilon}|\boldsymbol{v})$, from Equation (5.23):

$$\begin{aligned}\log Z(\boldsymbol{v}) &= \log Z(\boldsymbol{v}) \textstyle\int p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ d\boldsymbol{\Upsilon} \\ &= \textstyle\int \log Z(\boldsymbol{v})\ p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ d\boldsymbol{\Upsilon} \\ &= \textstyle\int \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})\ p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ d\boldsymbol{\Upsilon} - \int \log p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ d\boldsymbol{\Upsilon}.\end{aligned} \tag{5.52}$$

And its derivative with respect to v_e is:

$$\begin{aligned}\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e} &= \int (\frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e}\ p(\boldsymbol{\Upsilon}|\boldsymbol{v}) + \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})\ \frac{\partial p(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e})\ d\boldsymbol{\Upsilon} \\ &- \int (\frac{\partial \log p(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e}\ p(\boldsymbol{\Upsilon}|\boldsymbol{v}) + \log p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ \frac{\partial p(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e})\ d\boldsymbol{\Upsilon} \\ &= \int \frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e}\ p(\boldsymbol{\Upsilon}|\boldsymbol{v})\ d\boldsymbol{\Upsilon} \\ &=< \frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e} >_{p(\boldsymbol{\Upsilon}|\boldsymbol{v})} .\end{aligned} \tag{5.53}$$

Of particular interest is the expression $\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e} =< \frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e} >_{p(\boldsymbol{\Upsilon}|\boldsymbol{v})}$, since it is required for Equation (5.49). This integral can be estimated by sampling (Section 2.2.3). For our purpose, we employ Gibbs sampling with respect to the distribution $p(\boldsymbol{\Upsilon}|\boldsymbol{v})$. The generated samples are denoted by an additional subscript z. To derive $\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ we iteratively sample $\boldsymbol{\Psi}_z, \boldsymbol{\sigma}_z, \boldsymbol{y}_z, \boldsymbol{H}_z$ for N_z times, where the superscript ii denotes the specific sample, $ii = 1 : N_z$:

- **For** $p(\boldsymbol{\Psi}_z\ |\boldsymbol{v}, \boldsymbol{\sigma}_z, \boldsymbol{y}_z) = \mathcal{N}(\bar{\boldsymbol{\mu}}_{\Psi,z}, \bar{\boldsymbol{\Lambda}}_{\Psi,z}^{-1})$.
  The update Equations 5.35 -5.38 with $\boldsymbol{\sigma}_z, \boldsymbol{y}_z$ instead of $\boldsymbol{\sigma}, \boldsymbol{y}$ can be used:

  $$\Lambda_{\Psi,z,e} = v_e(\tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e} + \boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)}))^T \tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}), \tag{5.54}$$

  and

  $$\Lambda_{\Psi,z,e}\mu_{\Psi,z,e} = v_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})^T \tilde{\boldsymbol{D}}_e^T \boldsymbol{\sigma}_{z,e}^{(ii)}. \tag{5.55}$$

  Including the smoothing prior, the updated covariance and mean is:

  $$\bar{\boldsymbol{\Lambda}}_{\Psi,z} = \boldsymbol{\Lambda}_{\Psi,z} + (\boldsymbol{L}_\Psi^T \boldsymbol{H}_z^{(ii)} \boldsymbol{L}_\Psi), \tag{5.56}$$

  and

  $$\bar{\boldsymbol{\mu}}_{\Psi,z} \ = \bar{\boldsymbol{\Lambda}}_{\Psi,z}^{-1}(\boldsymbol{\Lambda}_{\Psi,z}\boldsymbol{\mu}_{\Psi,z}). \tag{5.57}$$

  $\boldsymbol{\Psi}_z^{(ii)}$ is sampled from $\mathcal{N}(\bar{\boldsymbol{\mu}}_{\Psi,z}, \bar{\boldsymbol{\Lambda}}_{\Psi,z}^{-1})$.

- **For** $p(\boldsymbol{H}_z\ |\boldsymbol{\Psi}_z) = \prod_{l=1}^{d_L} Gamma(a_{h,l,z}, b_{h,l,z})$
  The hyperparameters h_l, conditional on the other parameters follow a Gamma distribution $Gamma(a_{h,l}, b_{h,l})$ with

  $$a_{h,l,z} = a_{h,0} + \frac{1}{2}, \tag{5.58}$$

  $$b_{h,l,z} = b_{h,0} + \frac{1}{2}(\Psi_{z,l,1}^{(ii)} - \Psi_{z,l,2}^{(ii)})^2. \tag{5.59}$$

  $\boldsymbol{H}_z^{(ii)}$ is sampled from $\prod_{l=1}^{d_L} Gamma(a_{h,l,z}, b_{h,l,z})$.

- **For** $p(\boldsymbol{\sigma}_z \mid \boldsymbol{v}, \boldsymbol{\Psi}_z, \boldsymbol{y}_z) = \mathcal{N}(\bar{\boldsymbol{\mu}}_{\sigma,z}, \bar{\boldsymbol{\Lambda}}_{\sigma,z}^{-1})$
  The update Equations 5.41- 5.44 with $\boldsymbol{\Psi}_{z,i}, \boldsymbol{y}_{z,i}$ can be used:

$$\boldsymbol{\Lambda}_{\sigma,z,e} = v_e \boldsymbol{I}_{d_{S_e}}, \tag{5.60}$$

$$\boldsymbol{\Lambda}_{\sigma,z,e}\boldsymbol{\mu}_{\sigma,z,e} = v_e \Psi_{z,e}^{(ii)} \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}). \tag{5.61}$$

  Including the conservation law:

$$\bar{\boldsymbol{\Lambda}}_{\sigma,z} = k\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T + \boldsymbol{\Lambda}_{\sigma,z}, \tag{5.62}$$

$$\bar{\boldsymbol{\mu}}_{\sigma,z} = \bar{\boldsymbol{\Lambda}}_{\sigma,z}^{-1}(k\hat{\boldsymbol{B}}\boldsymbol{f} + \boldsymbol{\Lambda}_{\sigma,z}\boldsymbol{\mu}_{\sigma,z}). \tag{5.63}$$

  $\boldsymbol{\sigma}_{z,e}^{(ii)}$ is sampled from $\mathcal{N}(\bar{\boldsymbol{\mu}}_{\sigma,z}, \bar{\boldsymbol{\Lambda}}_{\sigma,z}^{-1})$.

- **For** $p(\boldsymbol{y}_z \mid \boldsymbol{\sigma}_z, \boldsymbol{\Psi}_z) = \mathcal{N}(\boldsymbol{\mu}_{y,z}, \boldsymbol{\Lambda}_{y,z}^{-1})$
  The update equations are the ones from Equations 5.45, 5.46 but without the contributions from the likelihood, i.e., without the terms dependent on τ:

$$\boldsymbol{\Lambda}_{y,z} = \sum_{e=1}^{d_{FE}} v_e \Psi_{z,e}^{(ii)} \Psi_{z,e}^{(ii)} \boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}} \boldsymbol{B}_{y,e} \boldsymbol{L}_{y,e}, \tag{5.64}$$

$$\boldsymbol{\mu}_{y,z} = \boldsymbol{\Lambda}_{y,z}^{-1} [\sum_{e=1}^{d_{FE}} v_e \Psi_{z,e}^{(ii)} \boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}^T (\boldsymbol{\sigma}_{z,e}^{(ii)} - \Psi_{z,e}^{(ii)} \tilde{\boldsymbol{D}} \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e})]. \tag{5.65}$$

  $\boldsymbol{y}_z^{(ii)}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}_{y,z}, \boldsymbol{\Lambda}_{y,z}^{-1})$.

Finally, we can write for the partition function derivative:

$$\begin{aligned} \frac{\partial \log Z(\boldsymbol{v})}{\partial v_e} &= < \frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e} >_{p(\boldsymbol{\Upsilon}|\boldsymbol{v})} \\ &= -\tfrac{1}{2} < ||\boldsymbol{\sigma}_{z,e} - \boldsymbol{D}_{z,e}(\boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e} + \boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e})||^2 >_{p(\boldsymbol{\Upsilon}|\boldsymbol{v})} \\ &= -\tfrac{1}{2}\tfrac{1}{N_z} \textstyle\sum_{ii=1}^{N_z} ||\boldsymbol{\sigma}_{z,e}^{(ii)} - \boldsymbol{D}_{z,e}^{(ii)}(\boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2 \\ &= -\tfrac{1}{2}\tfrac{1}{N_z} \textstyle\sum_{ii=1}^{N_z} ||\boldsymbol{\sigma}_{z,e}^{(ii)} - \Psi_{z,e}^{(ii)}\tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2, \end{aligned} \tag{5.66}$$

with N_z denoting the number of Gibbs samples that are used to calculate $< \frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e} >_{p(\boldsymbol{\Upsilon}|\boldsymbol{v})}$.

Algorithm 4 summarizes the fundamental steps of the resulting VB-EM algorithm. Steps $3-8$ correspond to the aforementioned VB-Expectation and $9-16$ to the VB-Maximization step. Currently, no specific convergence criteria are used in step 3 and 9, which is acceptable as long as the overall algorithm (convergence criterion in step

2) ensures convergence. For a convergence criterion in step 2 the lower bound $\mathcal{F}$ from Equation (5.47) cannot be used as $-\log Z(\boldsymbol{v})$ is not known. Therefore, we use

$$\hat{\mathcal{F}} = \mathcal{F} + \log Z(\boldsymbol{v}), \tag{5.67}$$

to study convergence.[3]

Algorithm 4 Algorithm for EM-based model error derivation

```
Initialize latent and model parameters, iter = 0
while F̂ has not converged do
    for i = 1 : N do
        update q(σ) using Equation (5.43), Equation (5.44)
        update q(H) using Equation (5.39), Equation (5.40)
        update q(Ψ) using Equation (5.37), Equation (5.38)
        update q(y) using Equation (5.45), Equation (5.46)
    end for
    for ii = 1 : N_z do
        sample σ_z^(ii) using Equation (5.60) - Equation (5.63)
        sample H_z^(ii) using Equation (5.58)- Equation (5.59)
        sample Ψ_z^(ii) using Equation (5.54) - Equation (5.57)
        sample y_z^(ii) using Equation (5.64) - Equation (5.65)
    end for
    update ∀v_e : ∂log Z(v)/∂v_e based on σ_z, Ψ_z, H_z, y_z and using Equation (5.66)
    update v using Equation (5.50)
    iter ← iter + 1
end while
```

5.2.2 Verification - Gibbs sampling

For validation purposes we derived and implemented the Expectation-Maximization algorithm using Gibbs sampling, instead of Variational Bayes. The resulting iterative optimization scheme is very similar. Only minor adjustments in the update equations are required. More information can be found in Appendix F.

5.3 Numerical illustration

In this section, we discuss examples from elastography and demonstrate the advantages and disadvantages of the proposed VB-EM algorithm for model calibration and

[3] It is not guaranteed that $\hat{\mathcal{F}}$ monotonically converges, as $\mathcal{F}$ does, however, this is not a requirement for a convergence study.

validation. The examples are based on the application of elastography, with the goal of estimating unknown material parameters based on displacement measurements. Besides the model parameters $\boldsymbol{\Psi}$ and possible hyper parameters, also the stresses $\boldsymbol{\sigma}$, displacements $\boldsymbol{y}$ and the precision of the model error $\boldsymbol{v}$ is of interest. We consider two set of examples from elastography. Example 1 provides insight into the characteristics of the algorithm. Furthermore, it is used for validation purposes and it is, for comparison reasons, also solved with Gibbs sampling. Example 2 shows that the algorithm can easily be applied to larger systems. An overview of the most important quantities/dimensions of the two examples is given in Table 5.1.

	Example 1	Example 2
Dimension of the observables $\hat{\boldsymbol{y}}$	840	5100
Dimension of the latent variables $\boldsymbol{y}$	840	5100
Dimension of the latent variables $\boldsymbol{\Psi}$	800	5000
Dimension of the latent variables $\boldsymbol{\sigma}$	2400	15000
Dimension of the model parameters $\boldsymbol{v}$	800	5000

Table 5.1: Summary of the dimensionalities of the observables, most important latent variables and model parameters for both discussed examples.

We assume an isotropic linear elastic material with a known Poisson ratio of $\nu = 0.45$ for all d_{FE} finite elements under the assumption of plane stress. The resulting dimension of the stresses and strains for a single triangular finite element is $d_{S_e} = 3$. The constitutive matrix for a single element e under plane stress is:

$$\boldsymbol{D}_e(\Psi_e) = \Psi_e \frac{1}{(1-\nu^2)} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{bmatrix},$$

which results in a constant $\tilde{\boldsymbol{D}}$ for all elements ($\boldsymbol{D}_e = \Psi_e \tilde{\boldsymbol{D}}_e$):

$$\tilde{\boldsymbol{D}}_e = \tilde{\boldsymbol{D}} = \frac{1}{(1-\nu^2)} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{bmatrix}.$$

Example 1 for VB-EM

We consider a two-dimensional domain $\Omega_0 = [0, L] \times [0, L]$ with $L = 20$ and $40 \times 20 = 800$ triangular finite elements. With a constant ψ within an element it results in a 800 dimensional vector $\boldsymbol{\Psi}$, $d_\Psi = 800$. With $d_{S_e} = 3$ constant stresses and strains within each element the vector of unknown stresses has the dimension $d_S = 2400$. We

assume Dirichlet boundary conditions in the normal components of the bottom and left boundary as well as Neumann conditions on the remaining boundaries (Figure 5.2)

$$\boldsymbol{u}_2 = \mathbf{0} \quad \text{on } x_1 = [0, L],\ x_2 = 0, \tag{5.68}$$

$$\boldsymbol{u}_1 = \mathbf{0} \quad \text{on } x_1 = 0,\ x_2 = [0, L], \tag{5.69}$$

and the following Neumann conditions on the remaining boundaries:

$$\begin{aligned} \hat{\boldsymbol{t}} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{on } x_1 \in [0, L],\ x_2 = L, \\ \hat{\boldsymbol{t}} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \text{on } x_1 = L,\ x_2 \in [0, L], \\ \hat{t}_1 &= 0, \qquad\qquad \text{on } x_1 \in [0, L],\ x_2 = 0, \\ \hat{t}_2 &= 0, \qquad\qquad \text{on } x_1 = 0,\ x_2 \in [0, L]. \end{aligned} \tag{5.70}$$

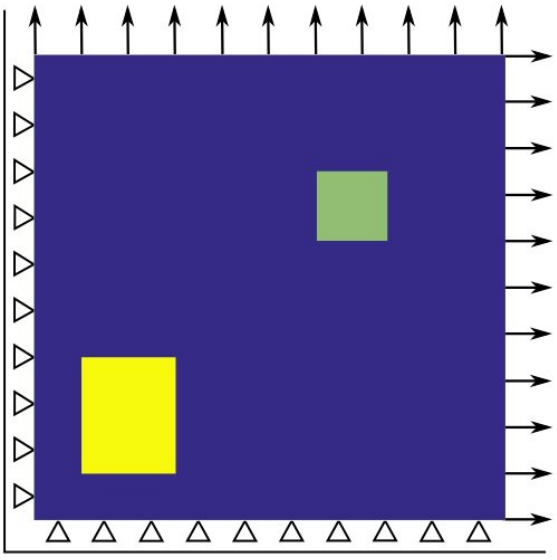


Figure 5.2: Configuration and reference. The blue and surrounding material is linear elastic with $\Psi = 1$, as well as the yellow inclusion with $\Psi = 0.2$. The green inclusion is based on a different material model, with its constitutive matrix given in Equation (5.71).

Synthetic data is obtained for a linear elastic material with $\Psi = 1$ in the surrounding tissue. The left bottom inclusion is also based on the linear elastic material model with $\Psi = 0.2$. The second inclusion (top right) is based on a different constitutive model

(anisotropic). The employed constitutive matrix $\tilde{\boldsymbol{D}}_{M,e}$ is:

$$\tilde{\boldsymbol{D}}_{M,e} = 6.2 \begin{bmatrix} 0.5 & -0.5 & 0.5 \\ -0.5 & 2 & 0.5 \\ 0.5 & 0.5 & 2 \end{bmatrix}. \tag{5.71}$$

Artificial measurements are generated with a $\mathrm{SNR} = 10^4$.

For solving the inverse problem, we assume that the material is linear, isotropic elastic with unknown elastic moduli. Similar to the material parameters $\boldsymbol{\Psi}$ (Equation (5.20)), we employ a smoothing prior for $\boldsymbol{v}$:

$$p_v(\boldsymbol{v}) \propto \exp^{-0.5\boldsymbol{v}^T(\boldsymbol{L}_\Psi^T \boldsymbol{H}_v \boldsymbol{L}_\Psi)\boldsymbol{v}}, \tag{5.72}$$

where $\boldsymbol{L}_\Psi$ is a difference operator and $\boldsymbol{H}_v = h_v \boldsymbol{I}_{d_L}$ a diagonal matrix. It accounts for the assumption that model errors are expected to be spatially correlated. In this example we employ $h_v = 0.1$.[4] We choose $N = 20$ and $N_z = 200$ for the numbers of iterations.[5] N_z is increased when the overall algorithm converges, to ensure that the convergence criteria is not be fulfilled based on a too strongly approximated gradient $< \frac{\partial \log \pi(\boldsymbol{\Upsilon}|\boldsymbol{v})}{\partial v_e} >_{p(\boldsymbol{\Upsilon}|\boldsymbol{v})}$.

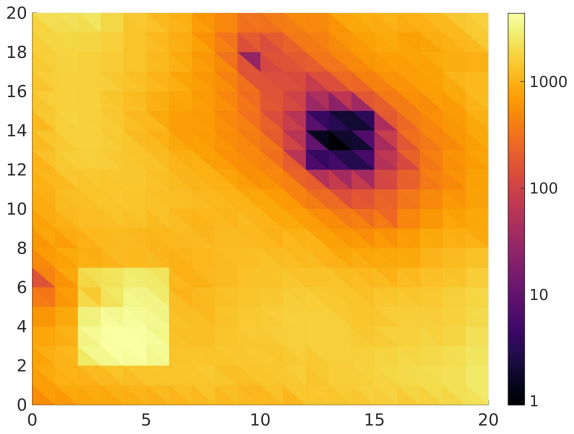


Figure 5.3: Point estimates of the precision $\boldsymbol{v}$ of the model error (colorbar in log-scale).

In Figure 5.3, the converged point estimates of the precision $\boldsymbol{v}$ of the model error is shown. We can see that the algorithm clearly identifies a significant model error. $\boldsymbol{v}$ is in this region two to four orders of magnitude smaller than in the rest of the problem.

[4] For the presented problem this seems to be an acceptable value, which is not too strict but still has a visible smoothing influence. A more sophisticated alternative, a hierarchical prior on h_v, is an alternative but has for the sake of simplicity not been employed.

[5] N refers to the number of iterations within a VB-E-step and N_z to the Gibbs iterations within a VB-M step (Algorithm 4). The larger number of N_z compared to N is chosen to derive an accurate expectation whereas for VB less iterations are necessary.

Despite the model inadequacy, the algorithm correctly identifies the material parameters $\mathbf{\Psi}$, see Figure 5.4[6].

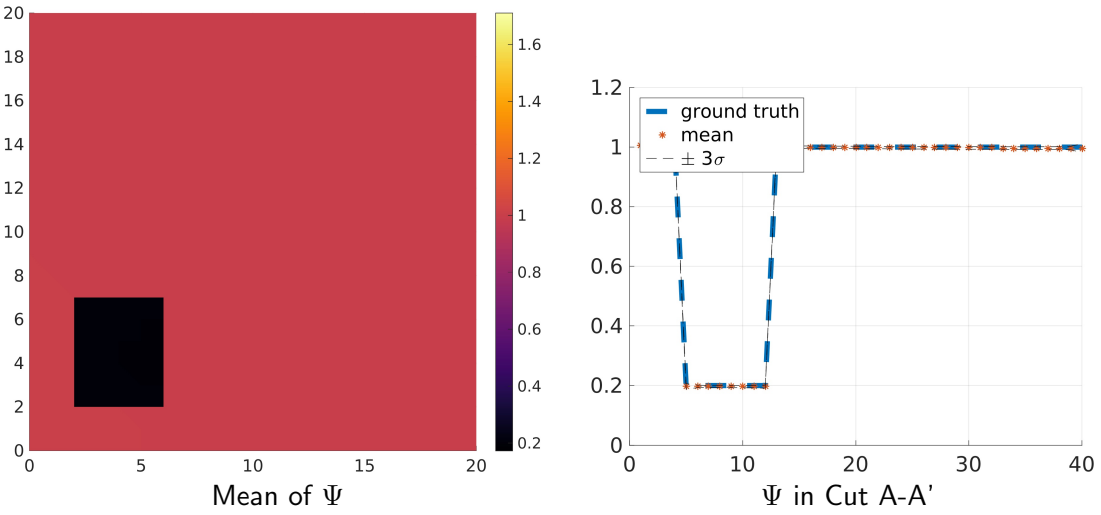


Figure 5.4: Left panel: Posterior mean of $\mathbf{\Psi}$. Right panel: Mean and posterior quantiles ($\pm 3\sigma$) of $\mathbf{\Psi}$ along the diagonal cut from (0, 0) to (20, 20).

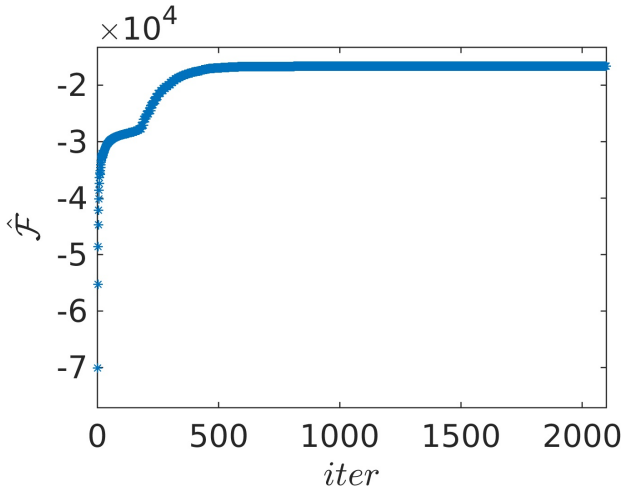


Figure 5.5: Evolution of $\hat{\mathcal{F}}$ over the number of $iter$-updates.

In Figure 5.6, the computed posterior characteristics of the stresses are shown. In particular, it is visible that the stresses in the whole domain are captured correctly. In addition, larger confidence intervals are visible in the region of the model error. This

[6]In the right figure the x-axis ranges from 1 to 40, as the diagonal cut goes through 40 triangular elements.

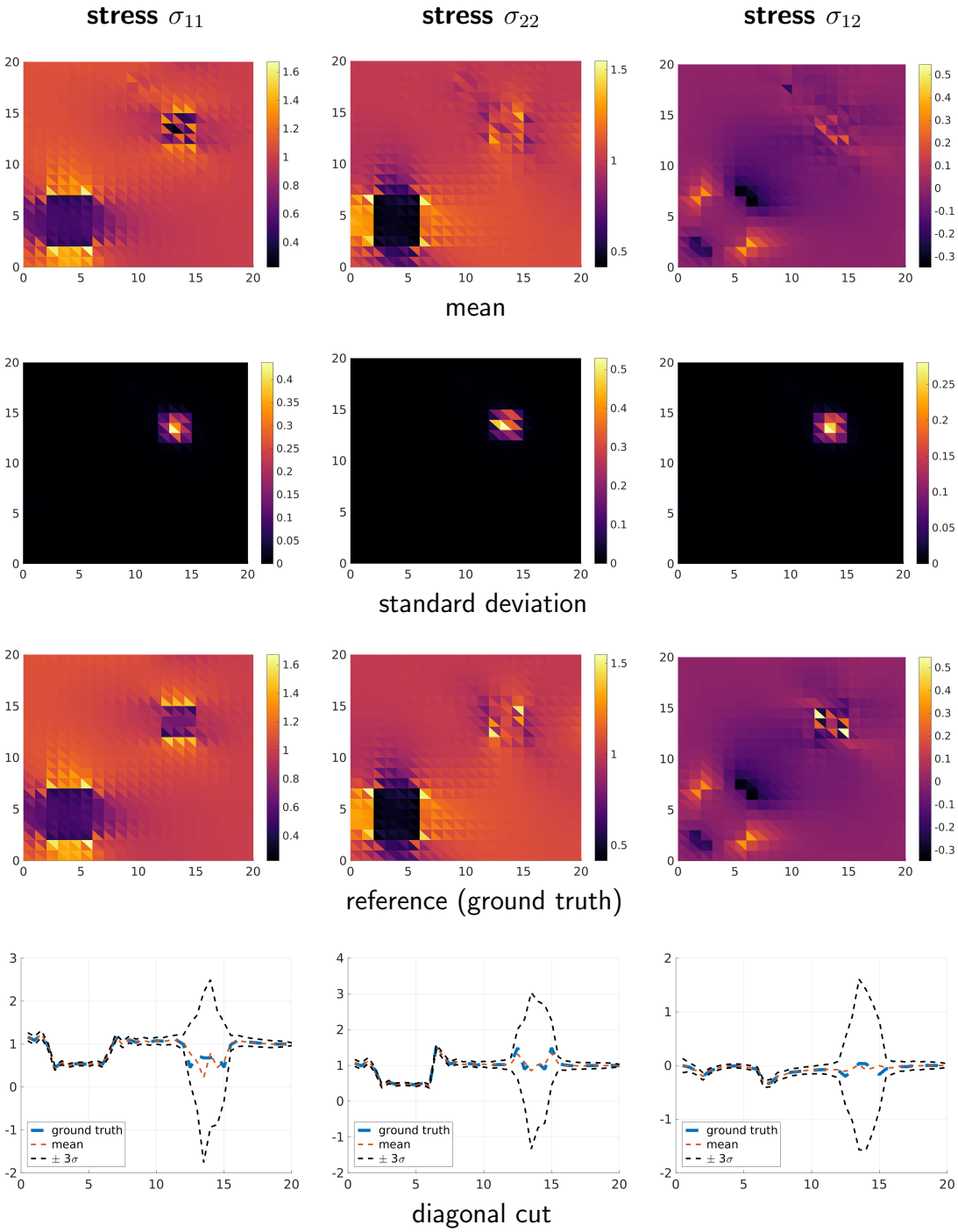


Figure 5.6: Comparison of the stresses' spatial distribution of the posterior mean and standard deviation with the ground truth. In the bottom row, the diagonal cuts from (0, 0) to (20, 20) of the posterior mean and credibility intervals are shown. The first column refers to σ_{11}, the middle column to σ_{22} and the right column to σ_{12} stresses.

shows that the algorithm can correlate the identified model error to resulting larger uncertainties in the stresses.

Figure 5.5 depicts the evolution of the lower bound $\hat{\mathcal{F}}$, from Equation (5.67). With regard to the required number of iterations to converge, we observe that the expectation $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ (Equation (5.66)) can accurately be approximated by $\frac{3}{2v_e}$. This relates to the derivative of the normalization term of $-\log p(\boldsymbol{y}, \boldsymbol{\sigma}_e, \Psi_e | v_e)$ from Equation (5.18) with respect to v_e. This approximation massively decreases the number of required iterations, and therefore, we use this approximation in the following. For more details we refer to Appendix G.

Example 2 for VB-EM

In the second example the problem is depicted in Figure 5.7. The domain is $\Omega_0 = [0, L] \times [0, L]$ with $L = 50$ and discretized with 100×50 finite elements resulting in the dimensions $d_\Psi = 5000$ and $d_S = 15000$. The same boundary conditions (Equation (5.68)- 5.70) and prior $p_v(\boldsymbol{v})$ (Equation (5.72)) from the previous example are employed. In this example, $h_v = 10$ and a $\mathrm{SNR} = 10^8$ is used.[7] In Figure 5.8, the

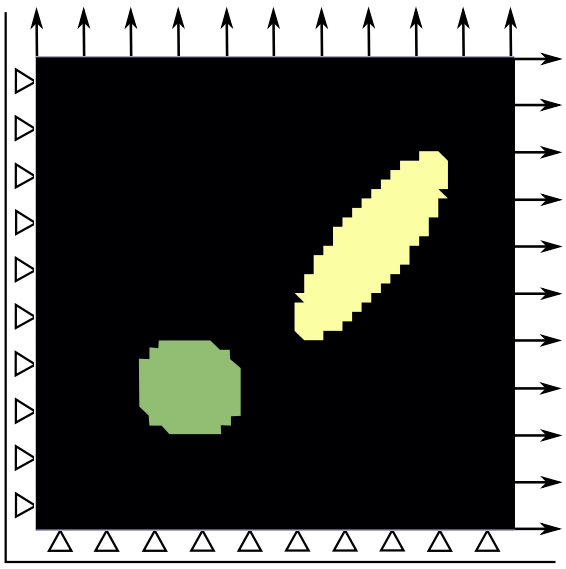


Figure 5.7: Configuration and reference. The black and surrounding material is linear elastic with $\Psi = 1$ and the yellow inclusion with $\Psi = 0.2$. The green inclusion is based on a different material model, its constitutive matrix is given in Equation (5.71).

converged point estimate of the precision of the model error $\boldsymbol{v}$ is shown. Also in this

[7] In contrast to the previous example, the model error is expressed with respect to the strains. Equation (5.16) is reformulated as: $(\boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}) = \boldsymbol{D}_e^{-1}\boldsymbol{\sigma}_e + \boldsymbol{c}_e$. The update equations need minor changes.

example, the algorithm clearly identifies the inclusion of the model error.

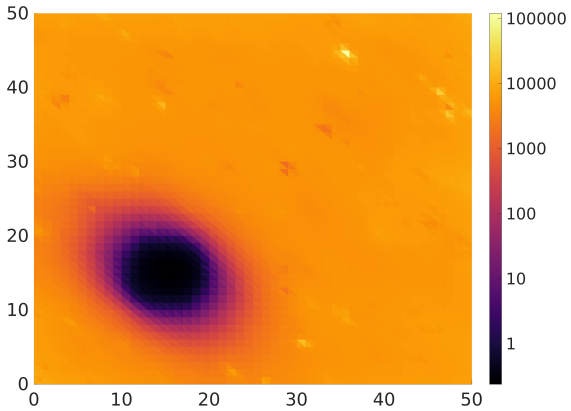


Figure 5.8: Point estimates of the precision $\boldsymbol{v}$ of the model error (colorbar in log-scale).

Also the derived material parameters $\boldsymbol{\Psi}$ (Figure 5.9) as well as the stresses (Figure 5.10) capture the truth correctly. Increased variances of the stresses are recognized in the region of the model error.

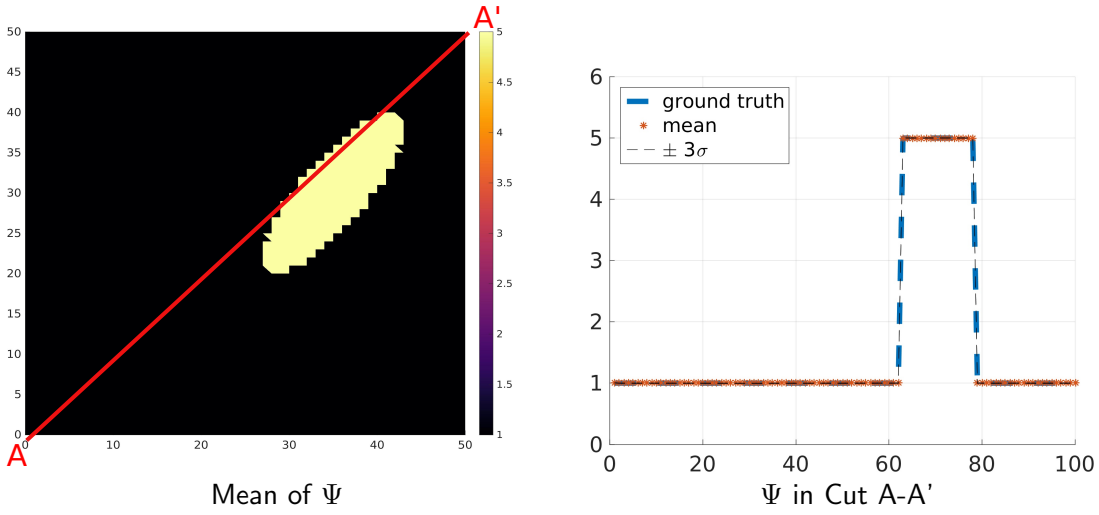


Figure 5.9: Posterior mean of $\boldsymbol{\Psi}$ and mean with posterior quantiles ($\pm 3\sigma$) along the diagonal cut from (0, 0) to (50, 50).

We also solved the same problem without quantifying model error. It can be seen that if $\boldsymbol{v}$ is not included, the material parameters and the stresses are not correctly identified, compare Figure 5.11 and Figure 5.12. This shows that the quantification of model error is important, as it could otherwise lead to erroneous conclusions and wrong interpretations. For example, one might think that there is a second inclusion with specific material parameters, e.g., in this example $\Psi \approx 16$.

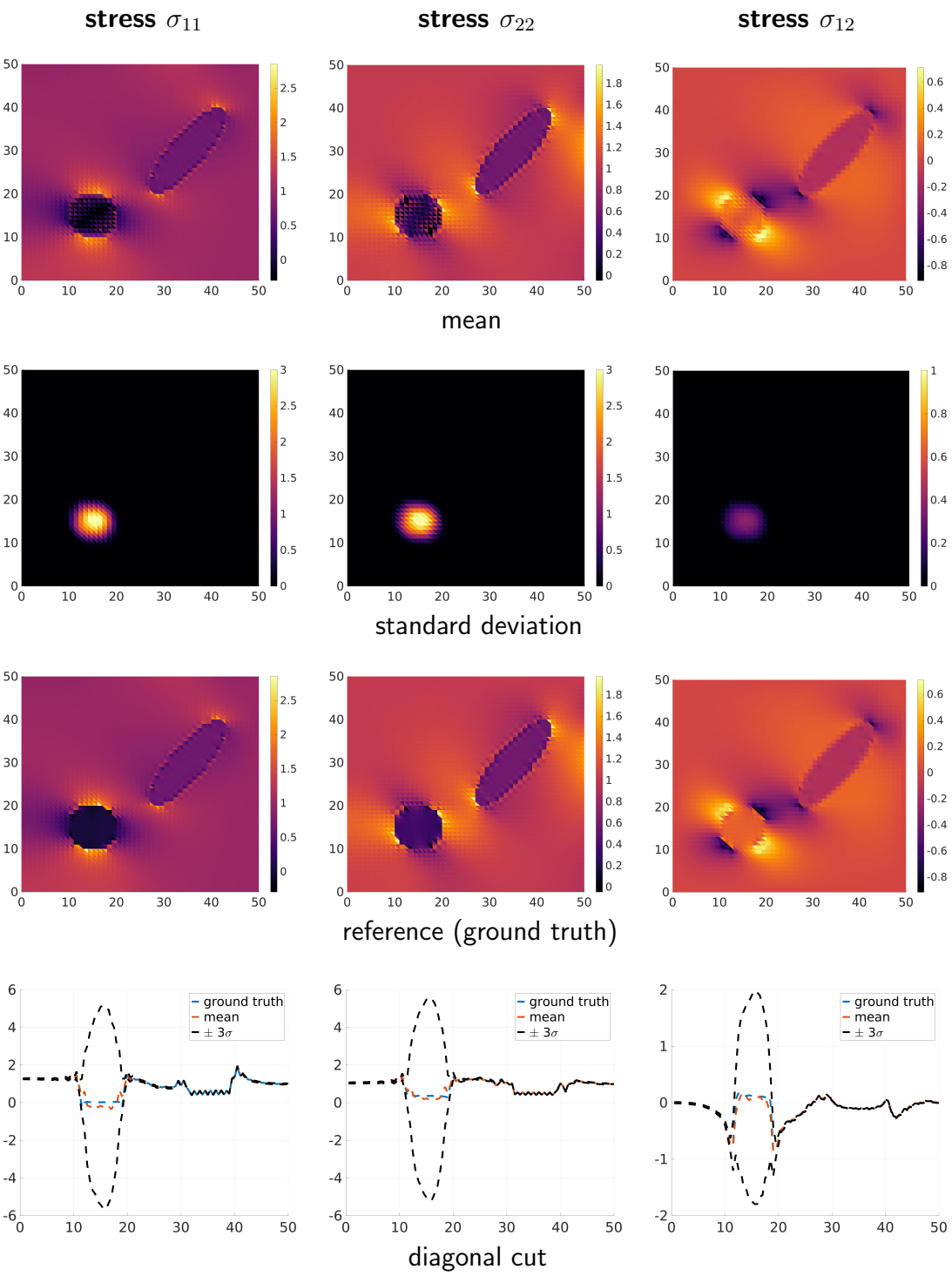


Figure 5.10: Comparison of the stresses' spatial distribution of the posterior mean and standard deviation with the ground truth. In the bottom row the diagonal cuts from (0, 0) to (50, 50) of the posterior mean and credibility intervals are shown. The first column refers to σ_{11}, the middle column to σ_{22} and on the right to σ_{12} stresses.

The omitted model error influences the overall domain. Also in regions where there is no model error incorrect material parameters and stresses are determined. Without a quantification of the model error, the quantified variances in the stresses are close to zero.

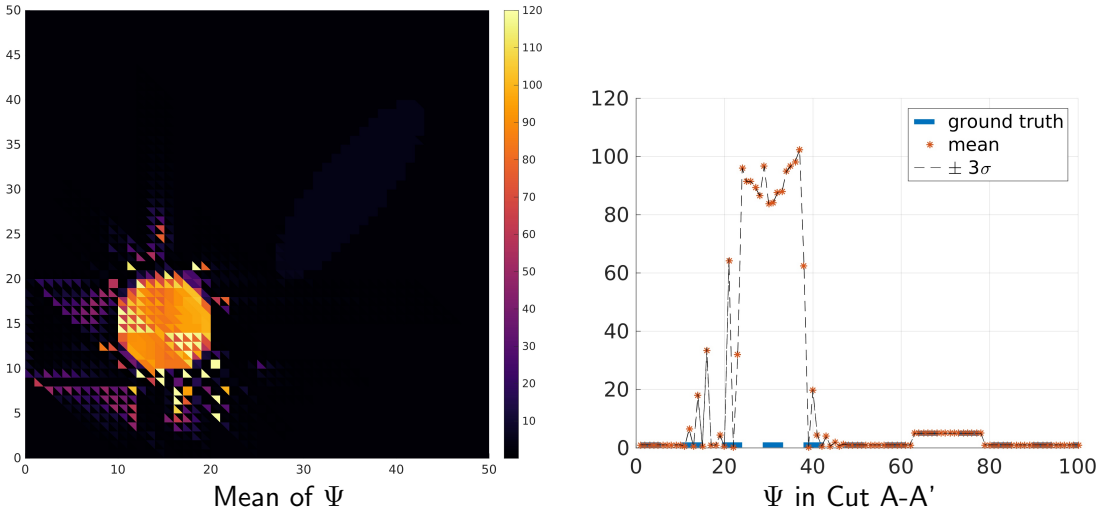


Figure 5.11: Left: Posterior mean of $\boldsymbol{\Psi}$. Right: Posterior statistics of $\boldsymbol{\Psi}$ along the diagonal cut from (0, 0) to (50, 50), derived without a quantification of model error.

We want to stress that this example has very small measurement noise ($\mathrm{SNR} = 10^8$). In examples with larger measurement noise, the fluctuations in the mean of the material parameters and stresses are smaller. This is balanced by the smoothing prior of $\boldsymbol{\Psi}$. However, also in examples with larger noise levels, an accurate quantification of the latent variables without quantifying the model error was not possible for the discussed examples. However, if a system can correctly be described without quantifying model error, it will prefer to do so and only quantify model error if it is necessary.

Example 1 for Gibbs-EM

For validation purposes we performed Gibbs sampling (Section 5.2.2) in order to assess the overall accuracy of the approximation for the first and smaller example (Figure 5.2). In Figure 5.13, the converged point estimate of the precision of the model error $\boldsymbol{v}$ is shown. As beforehand with VB, the algorithm could correctly identify the inclusion of model error. However, comparing the derived point estimates with those obtained with VB (Figure 5.3), differences are visible: The precision of the model error obtained with VB is, in the region without model error, two orders of magnitude smaller than that obtained with Gibbs sampling. By looking at the individual terms, we could observe that this discrepancy correlates to differences in the expectations of Ξ_e, derived by

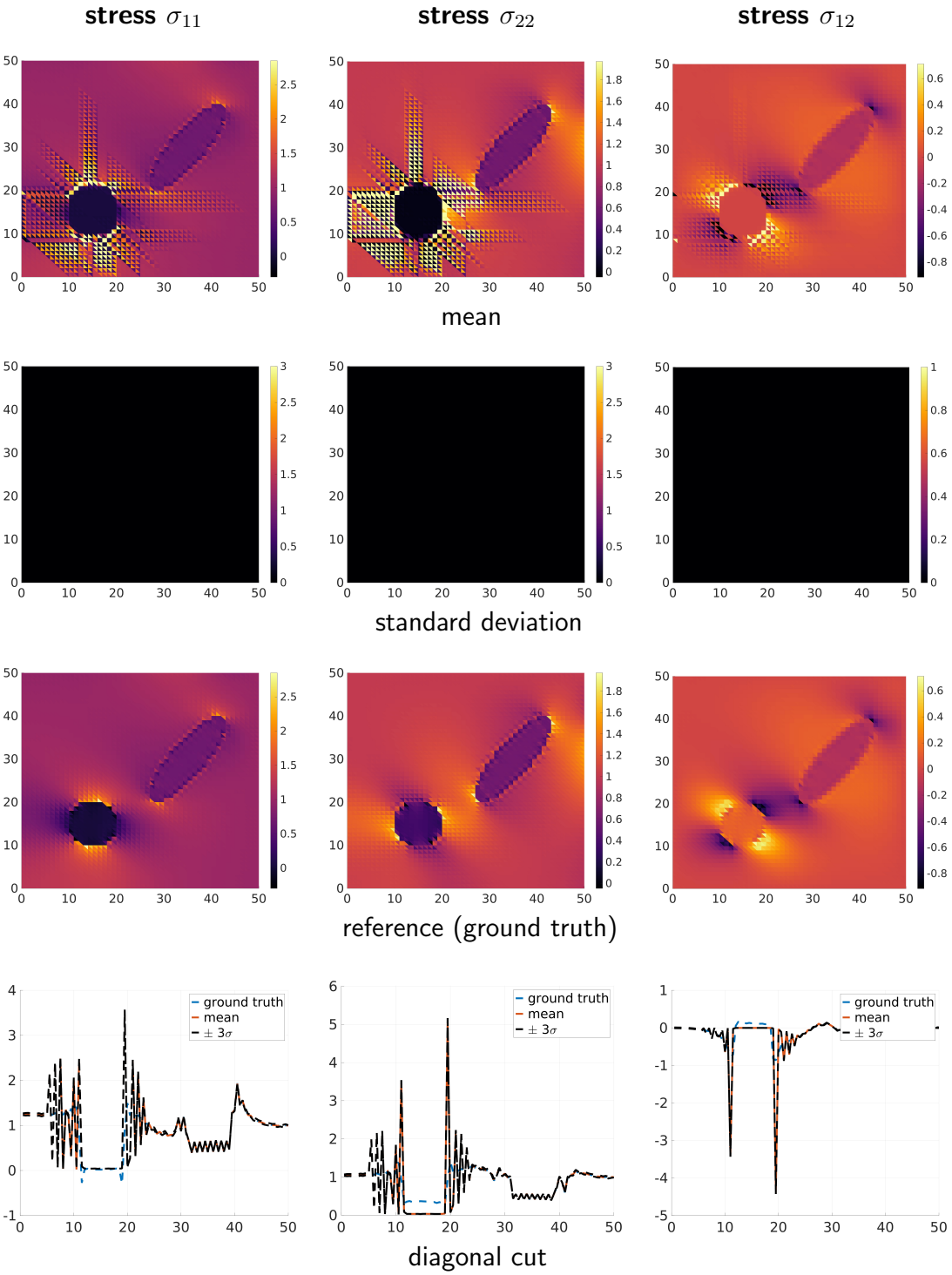


Figure 5.12: Comparison of the stresses' spatial distribution of the posterior mean and standard deviation with the ground truth. In the bottom row the diagonal cuts from (0, 0) to (50, 50) of the posterior mean and credibility intervals are shown. The first column refers to σ_{11}, the middle column to σ_{22} and on the right to σ_{12} stresses. The results are derived without a quantification of model error.

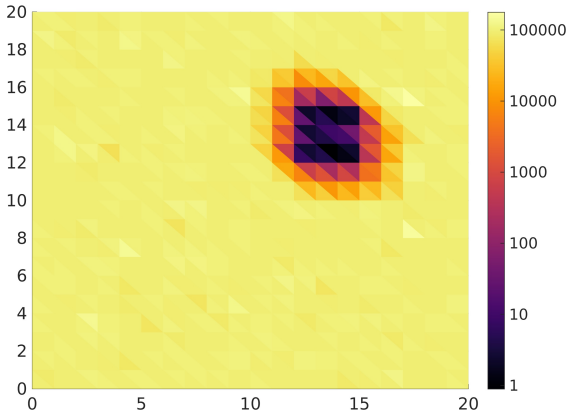


Figure 5.13: Point estimates of the precision $\boldsymbol{v}$ of the model error.

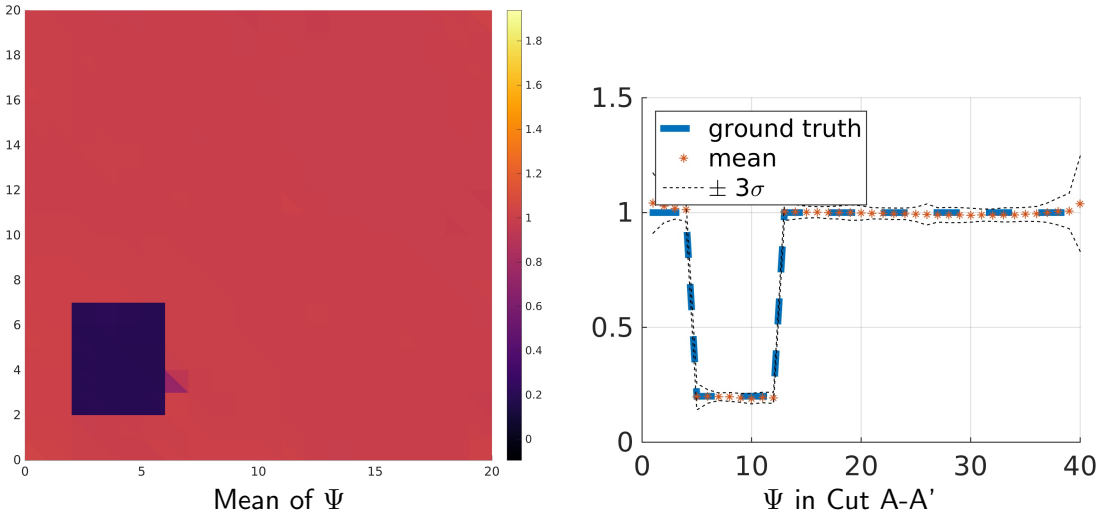


Figure 5.14: Posterior mean of $\boldsymbol{\Psi}$ and mean with posterior quantiles ($\pm 3\sigma$) along the diagonal cut from (0, 0) to (20, 20), obtained with Gibbs sampling in the E-step.

VB (Equation (5.51)) compared to Gibbs sampling (Equation (F.13)). Even though in both cases the difference between the stresses and the model predicted stresses should be negligible for elements without model error, in the calculations with VB the variances $\bar{\boldsymbol{\Sigma}}_{\sigma,e}, \bar{\Sigma}_{\Psi,e}, \boldsymbol{\Sigma}_{u,e}$ produce a disagreement. Although in the calculation with Gibbs sampling the samples also vary and show some variations, the latent variables are also correlated (Figure 5.16). This results in a smaller Ξ_e, compared to VB. For VB, this correlation is neglected by the mean-field-approximation, assuming independent latent variables.

In Figure 5.16, for some selected elements e the posterior distribution and the correlation of the stresses with the material parameter is shown. Elements belonging to the

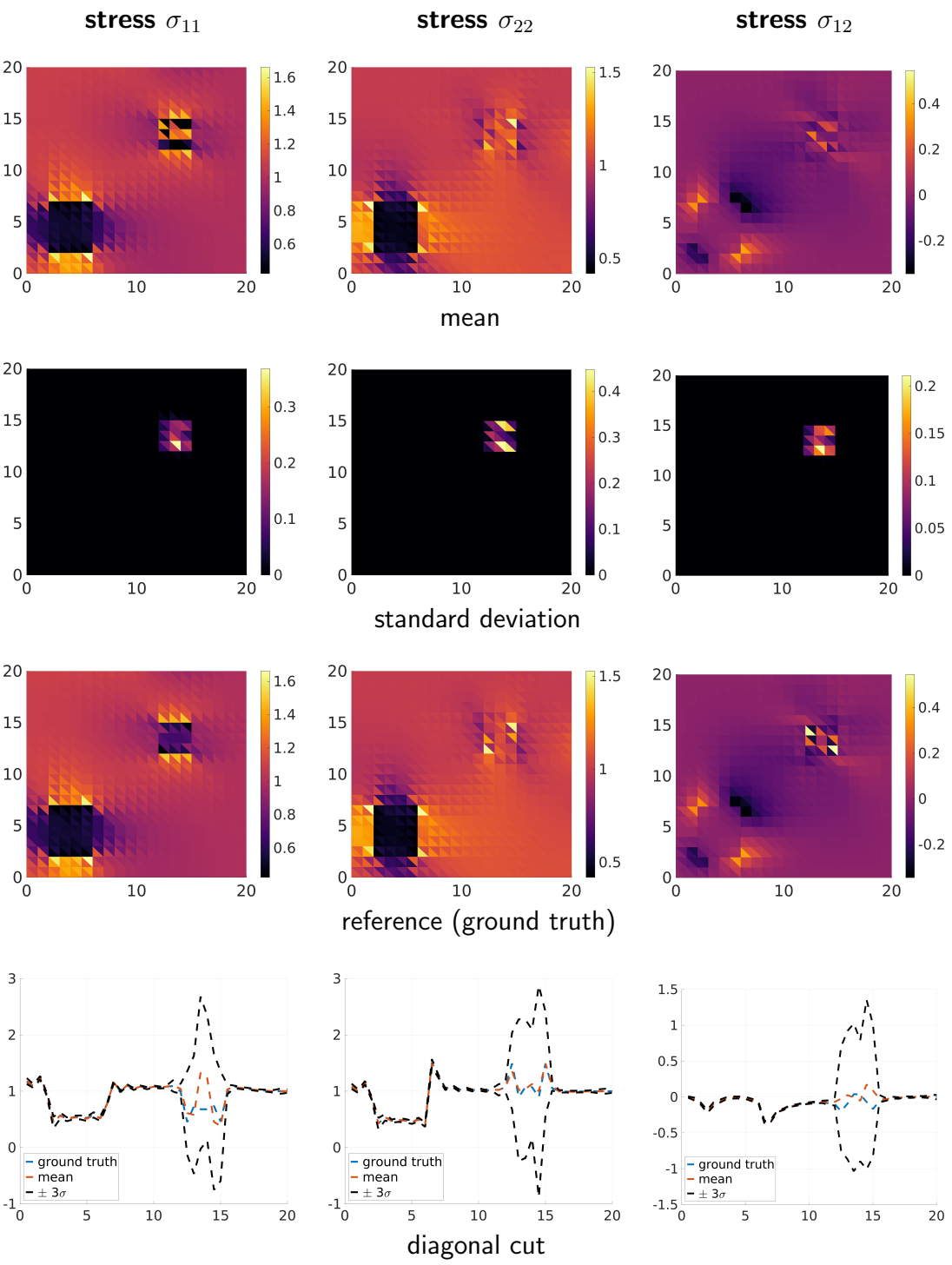


Figure 5.15: Comparison of the stresses' spatial distribution of the posterior mean and standard deviation with the ground truth. In the bottom row the diagonal cuts from (0, 0) to (20, 20) of the posterior mean and credibility intervals are shown. The first column refers to σ_{11}, the middle column to σ_{22} and on the right to σ_{12} stresses. These results are obtained with Gibbs sampling for the E-step.

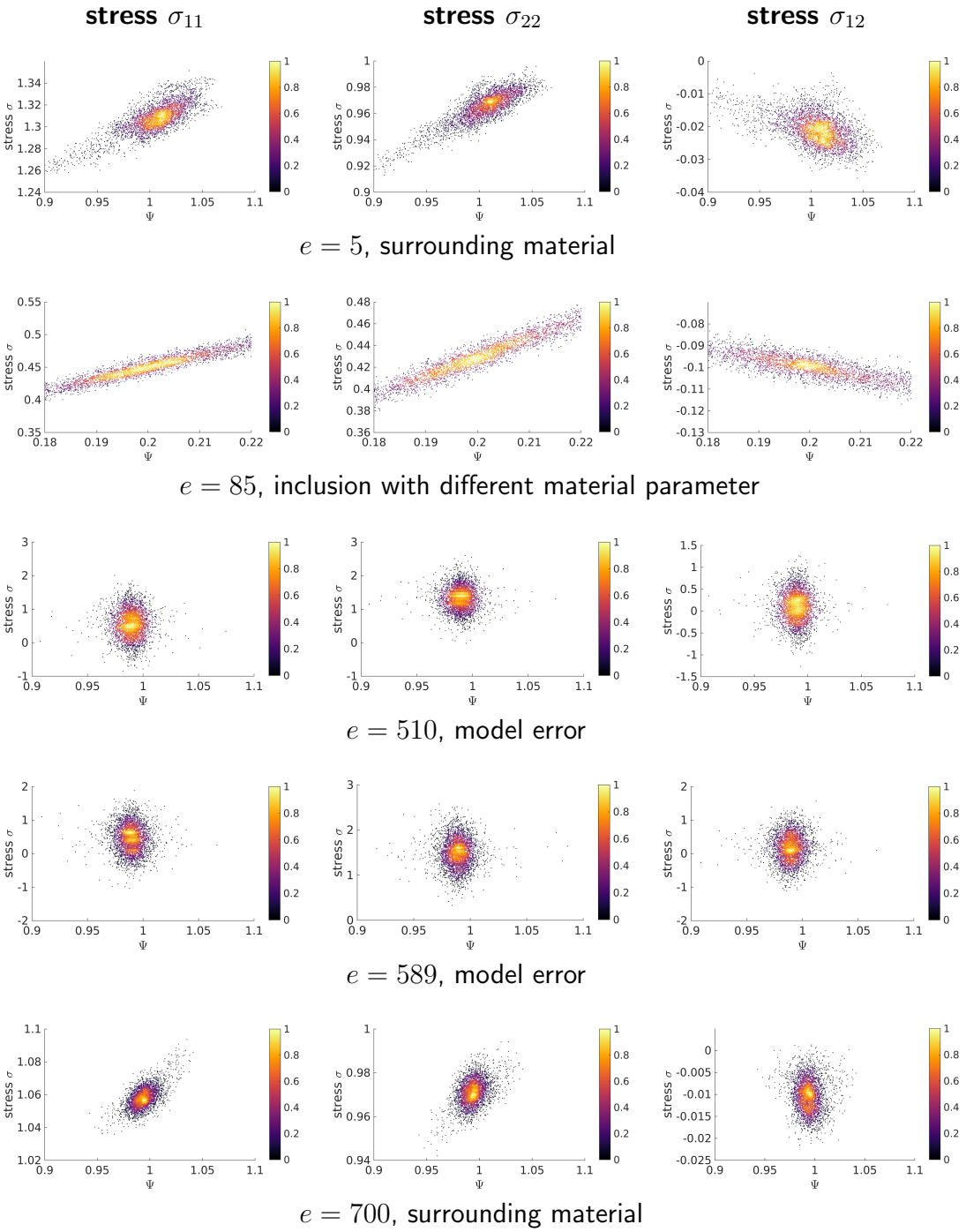


Figure 5.16: Two-dimensional plot of the posterior samples for specified elements e. In the first column the correlation of the stresses $\sigma_{11,e}$ to Ψ_e is shown. The same for the stresses $\sigma_{22,e}$ and $\sigma_{12,e}$ in column two and three.

surrounding tissue show some correlation between the stresses and the material parameter for a single element (e.g., $e = [5, 700]$). This correlation increases for elements in the inclusion with a different true material parameters ($e = 85$). In contrast to that, for elements where there is model error this correlation is not visible ($e = [510, 589]$). For elements with model error, VB and Gibbs sampling identify the same point estimates for v_e (see Figure 5.3 and Figure 5.13). This can be explained by the fact that without any correlation within the latent variables VB can correctly approximate the posterior. Since the model error is correctly recognized, the latent variables such as stresses, displacements and material parameters have higher variances and less correlation within each other in those elements. In contrast to that, for elements with no model error and a stronger correlation within the latent variables, VB does not correctly identify the precision of the model error v_e, as discussed in the previous paragraph.

In Figure 5.14, the posterior mean and the posterior statistics of $\mathbf{\Psi}$, along the diagonal cut from (0, 0) to (20, 20), are shown. Comparing these statistics with the ones obtained with VB (Figure 5.4), differences in the variances are visible. As there is also a measurement error with a resulting $SNR = 10^4$, we also expect variations for elements without model error as shown in Figure 5.14. Finally, in Figure 5.15, the derived posterior statistics of the stresses are shown. Besides minor fluctuations they coincide to the results obtained with VB (Figure 5.6).

5.4 Summary

Most Bayesian strategies are deficient in quantifying model inadequacy. Therefore, in most studies it is implicitly assumed that the model is perfect. In this chapter, we introduced an effective quantification of model fidelity with an intrusive Bayesian framework to quantify constitutive model error. Based on recent work [108], we opened the classical black-box forward problem to assess the model fidelity in a physical context. In addition, we added a consistent normalization term, which allows for more flexible prior assumptions. We combined various model equations as well as further a priori information in a joint prior. Moreover, a Variational Bayesian Expectation Maximization scheme was used for computational efficiency.

For the employed examples from elastography, we showed that the material parameters, as well as model inadequacies, could be quantified. This is particularly important as the material parameters for an assumed material model are used for non-invasive medical diagnosis. Thus, identified model inadequacies are helpful for three reasons: Firstly, for knowing that the inferred material parameters are not trustworthy and therefore should not be used for diagnosis. Secondly, identified model inadequacy can be used as an indicator for a suspicious tissue. As biological tissues can also be identified by their material model (compare Chapter 1 and specifically the degree of nonlinearity of cancerous tissue), an identified model error, referring to an underlying different model,

can be used as an indicator for a distrustful and maybe cancerous tissue. Thirdly, deciding when an alternative model, with a different constitutive relation, needs to be used.

The algorithm infers, apart from model parameters and model fidelity, additional latent variables, e.g., stresses. Although the number of (auxiliary) variables increases, compared to a classical black-box forward problem, it has a computational advantage. Namely, that no forward problem needs to be solved, which is usually computationally expensive. For future work, we propose to exploit this advantage in more detail and to push the computational cost to new lower limits: As no forward model needs to be solved, advanced, element-wise or block-wise solution strategies can be exploited.

For validation purposes, we also employed Gibbs sampling. It was shown that, although the overall results coincide (e.g., mean values of the latent variables), differences were observed in the credibility intervals. This can be attributed to the mean-field-assumption, which assumes that stresses, displacements and material parameters are independent. Therefore, an interesting extension of the current work would be to effectively quantify (local) correlations, e.g., by a band covariance matrix to attenuate the mean field assumption.

Another possible extension of the developed framework is to expand it to nonlinear material models. However, as the stress-strain relation is dependent on the strains more advanced schemes become necessary. For example, inspired by existing approaches for nonlinear finite element methods (Section 2.3.3), iterative models may be employed.

Chapter 6

Discussion, Summary and Outlook

> *There are no routine statistical questions, only questionable statistical routines.*
>
> Sir David R. Cox, 1924-today.

Summary

This thesis has successfully developed, implemented and verified a novel Bayesian framework for large-scale nonlinear inverse problems. It has shown how limitations of current algorithms, i.e., computationally-demanding forward models, the 'curse of dimensionality', capturing multimodality and quantifying model error can be addressed. We have developed a novel Variational Bayesian (VB) framework and utilizing its conceptual advantages (e.g., solving a computationally tractable optimization problem), we demonstrated its ability to efficiently perform variational inference to provide approximate inference. A novel dimensionality reduction method has been developed based on a fully Bayesian formulation. The method identifies lower dimensional subspaces where most of the posterior variance is concentrated. In order to adaptively identify the cardinality of the reduced coordinates an information-theoretic criterion has been proposed. The posterior approximations have been obtained with a limited number of calls to the computationally expensive forward solver (in the considered examples fewer than 35 forward calls for 2500 material parameters).

In addition, we have successfully extended the Variational Bayesian algorithm and the novel dimensionality reduction to multimodal distributions by using a mixture of Gaussians. More specifically, each of the mixture components can identify different lower-dimensional substructures. This can be compared with existing frameworks that

also use VB with a mixture of Gaussians. Those frameworks are limited to very low-dimensional examples with up to three latent variables and require data from multiple experiments [93, 79, 94]. Recently, this approach has been extended to a problem with six latent variables, as many as two mixture components and data from a single experiment [109]. In contrast, we have shown that the developed framework with the novel dimensionality reduction can be used to effectively and accurately quantify multimodal posteriors for high-dimensional problems. This has been demonstrated on examples with 2500 unknown latent material parameters and with data based on a single experiment. The algorithms have been verified by importance sampling, which showed that the approximations are trustworthy and the bias introduced by the approximations is small and can efficiently be corrected (i.e., with very few forward model calls).

The quantification of constitutive model inadequacy has also been investigated in this thesis. By opening the classical black-box forward problem, model fidelity can be assessed in a physical context. Our approach is based on recent work [108] that develops an intrusive algorithm by embedding the quantification of the model error in a selected, phenomenological submodel. However, we included a consistent normalization term within the framework, which allows for more flexible prior assumptions. In addition, in comparison to the work in [100] that was limited to five latent variables, we solved high-dimensional problems with more than 25000 unknowns. Intrusive approaches avoid several drawbacks of non-intrusive methods [46, 98], such as entanglement with measurement noise, possible violations of physical constraints and missing predictive capabilities [34]. The developed framework can also incorporate a VB approximation, which has been compared, in this thesis, to outcomes of empirical methods. The comparison identified an important correlation between the latent variables that is neglected by the structured mean-field approximation. This directs to potential future research, which we will discuss presently, in the outlook.

Finally, we have shown that uncertainty quantification can also successfully be applied for large-scale engineering applications. The framework developed in this thesis is generally applicable but in this thesis has been specifically applied to static, nonlinear elastography. For this application, the importance of statistical inference was shown. A lack of knowledge regarding the underlying uncertainties results in inaccurate information and overconfident conclusions. This is particularly misleading when the posterior distribution is not unimodal and sharply peaked, e.g., a multimodal probability distribution. The analyst or medical practitioner can use this information to draw a precise, patient-specific and non-invasive diagnosis. The importance of probabilistic elastography is particularly manifested in the boundaries of the inclusions (tumors) [210, 211]. The inclusions and their specific effect on the diagnostic results can be better classified with statistical inference.

Outlook

By solving some of the main issues of statistical inference for high-dimensional problems, we have opened up some additional interesting investigations, questions and ideas. These may serve as possible and prosperous directions of future work with the potential to impact multiple applications:

- The examples in this thesis are based on nonlinear (quasi-)static elastography. An interesting next step would be to include *viscoelastic* materials. It has been found that pathological changes, e.g., in cancerous breast tumors also affect the viscous properties, in particular the attenuation of the tissue [224]. To solve a linear viscoelastic elastography problem, the forward problem can be formulated with Fourier transformations in the frequency domain for steady-state (time-harmonic) vibrations [225]. Then the derived probabilistic methods do not need to be adjusted. However, as soon as nonlinear viscoelastic material models are used, or a non-harmonic deformation is applied, the system becomes more complex and additional adjustments are required [226].

- Within this thesis, the measurement noise is assumed to be white Gaussian noise. This is a valid assumption based on the central limit theorem, as long as the number of different sources of measurement errors is large enough [85]. Measurement errors can have several origins, e.g., experimental errors such as human error, mistakes in the data entries, or fault in the design of the experiment. However, in strain-based elastography the (ultrasound) images are registered to derive the deformation map. This may introduce not only significant but also systematic errors that do not coincide with the assumption of Gaussian measurement noise. Salt-and-pepper noises are common in images as well [170]. The issue of incorrectly identified measurement errors can be resolved by extending the research in the thesis as follows:

  - Adjustments of the prior assumptions, e.g., block prior probabilities for salt-and-pepper-noise [227].
  - Quantifying the uncertainties in the (ultrasound) image itself [228].
  - Quantifying the uncertainties within the image registration process [110, 111, 114, 112, 113].
  - Reduction of the measurement errors by circumventing an image registration and by using the images directly as measurements [229].

- In the previous chapter the quantification of model inadequacy in an intrusive framework was been developed. This can be used as a foundation for various future research directions:

  - By quantifying the model error with Variational Bayes, we have observed that the outcome using the mean-field approximation of independent latent variables is ambitious. An interesting extension of the current work would be to effectively quantify local correlations, e.g., by a band covariance matrix to attenuate the mean field assumption.
  - The big advantage of not solving a forward problem should be exploited in more detail. An advanced, element-wise, local solution strategy decreases the scaling of computational cost with regard to the number of elements. This results in an effective algorithm for high-dimensional inverse problems.
  - The two main approaches in the thesis could be combined. This would lead to a model error quantification as well as a dimensionality reduction for each group of latent variables (displacements, stresses and material parameters). Their lower-dimensional subspaces could be identified by the proposed, incremental, iterative algorithm.

- On a medium-term, it is also essential to verify and validate the presented methods with experimental data and clinical studies. For a *verification*, the specification and quality of the novel elastography approaches need to be evaluated. Although several aspects of clinically relevant problems are already captured in the investigated numerical methods, working with real data may add additional difficulties for which adjustments might be necessary. The measurement errors may be non-Gaussian and have strong outliers or systematic behavior. Alternatively, the assumed boundary conditions may be incorrect. Besides verifying the methods, validation is also of interest. To *validate* possible applications, it is important to adapt results for the applicant such that the information of the outcome is in accordance with the application. For instance, when showing the results of elastography, ambiguous results should be avoided.

The presented suggestions are only a few of many potential, future research opportunities that are directed to the field of Bayesian inference and its application in elastography. An even more important extension of this thesis are the many practical applications of the developed framework (see introduction). The results of this thesis hopefully broaden the scientific field and encourage further research on Bayesian variational inference applications.

On a final note, the application of elastography is also a good example to show that not only the way of answering but also of posing (engineering) questions is important. In my opinion, questions and answers are often posed too simply. Instead of asking and answering questions as a dichotomist, for which only two possible answers, namely 'yes' and 'no' exist, more specific questions are advisable. In real life, a measure of

uncertainty reflects a finer nuanciated attitude of life and its complexities. Not only the way of how to solve problems, but particularly the way of how to pose a question needs to be adjusted in my mind. Instead of asking 'Do I have cancer?' when ultrasound images are taken, it should be 'What is the probability for me having cancer?'. Although the question itself barely changes, it describes a different attitude. A patient can react to the second question in a significantly more nuanced form. If the answer is 'the percentage of not having cancer is 90%', the patient has room for deciding if he or she wants actions to follow the 90% or the 10% regime. Possible actions when focusing on the 10% regime could be further investigations, such as a biopsy, more medical examinations or a therapy which otherwise would be missed. Such an involvement of the patient in the decision making process would not be possible when the answer is simply 'yes, you are healthy and you have no cancer'. In this context, elastography with uncertainty quantification helps to give *more accurate uncertain answers*.

Appendix A

Expectation-maximization for the μ-prior

Due to the analytical unavailability of $\log p_\mu(\boldsymbol{\mu})$ and its derivatives $\frac{\partial \log p_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ we employ an Expectation-Maximization scheme which we describe in here for completeness [187, 135]. Proceeding as in Equation (3.8), i.e., by making use of Jensen's inequality and an arbitrary distribution $q(\boldsymbol{\Xi})$ we can bound $\log p_\mu(\boldsymbol{\mu})$ as follows:

$$\begin{aligned}\log p_\mu(\boldsymbol{\mu}) &= \log \int p(\boldsymbol{\mu}|\boldsymbol{\Xi})\ p(\boldsymbol{\Xi})\ d\boldsymbol{\Xi} \\ &= \log \int \tfrac{p(\boldsymbol{\mu}|\boldsymbol{\Xi})p(\boldsymbol{\Xi})}{q(\boldsymbol{\Xi})} q(\boldsymbol{\Xi})\ d\boldsymbol{\Xi} \\ &\geq \int q(\boldsymbol{\Xi}) \log \tfrac{p(\boldsymbol{\mu}|\boldsymbol{\Xi})p(\boldsymbol{\Xi})}{q(\boldsymbol{\Xi})}\ d\boldsymbol{\Xi} \\ &=< \log p(\boldsymbol{\mu}|\boldsymbol{\Xi}) >_{q(\boldsymbol{\Xi})} + < \log \tfrac{p(\boldsymbol{\Xi})}{q(\boldsymbol{\Xi})} >_{q(\boldsymbol{\Xi})} .\end{aligned} \tag{A.1}$$

The inequality above becomes an equality only when $q(\boldsymbol{\Xi}) \equiv p(\boldsymbol{\Xi}|\boldsymbol{\mu})$, i.e., it is the actual posterior on $\boldsymbol{\Xi}$ given $\boldsymbol{\mu}$. The latter can be readily established from Equations (3.21) and (3.23) based on which $p(\boldsymbol{\Xi}|\boldsymbol{\mu}) = \prod_{m=1}^{d_L} Gamma(a_{\xi_m}, b_{\xi_m})$ where:

$$a_{\xi_m} = a_\xi + \frac{1}{2}, \quad b_{\xi_m} = b_\xi + \frac{1}{2}(\mu_{k_m} - \mu_{l_m})^2. \tag{A.2}$$

This suggests a two-step procedure for computing $\log p_\mu(\boldsymbol{\mu})$ and $\frac{\partial \log p_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ for each $\boldsymbol{\mu}$:

(E-step) Find $p(\boldsymbol{\Xi}|\boldsymbol{\mu}) = \prod_{m=1}^{d_L} Gamma(a_{\xi_m}, b_{\xi_m})$ from Equation (A.2).

(M-step) Find $\log p_\mu(\boldsymbol{\mu})$ and $\frac{\partial \log p_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ from Equation (A.1) for $q(\boldsymbol{\Xi}) \equiv p(\boldsymbol{\Xi}|\boldsymbol{\mu})$ as follows:

$$\begin{aligned}\log p_\mu(\boldsymbol{\mu}) &=< \log p(\boldsymbol{\mu}|\boldsymbol{\Xi}) >_{q(\boldsymbol{\Xi})} = -\tfrac{1}{2}\boldsymbol{\mu}^T \boldsymbol{L}^T < \boldsymbol{\Xi} > \boldsymbol{L}\boldsymbol{\mu} \\ \tfrac{\partial \log p_\mu(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} &= \tfrac{\partial}{\partial \boldsymbol{\mu}} < \log p(\boldsymbol{\mu}|\boldsymbol{\Xi}) >_{q(\boldsymbol{\Xi})} \\ &=< \tfrac{\partial}{\partial \boldsymbol{\mu}} \log p(\boldsymbol{\mu}|\boldsymbol{\Xi}) >_{q(\boldsymbol{\Xi})} \\ &= -\boldsymbol{L}^T < \boldsymbol{\Xi} > \boldsymbol{L}\boldsymbol{\mu},\end{aligned} \tag{A.3}$$

where $< \mathbf{\Xi} >=< diag(\xi_m) >_{q(\mathbf{\Xi})}= diag(\frac{a_{\xi_m}}{b_{\xi_m}})$.

Appendix B

Variational lower bound for MoG

The lower bound from Equation (4.23) combined with the optimal probability distributions q^{opt}, Equation (4.24), is:

$$
\begin{aligned}
\hat{\mathcal{F}}(q^{opt}(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s),\boldsymbol{T}) \quad & = \tfrac{d_y}{2} < \log\tau >_\tau && (< \log p(\hat{\boldsymbol{y}}|s,\boldsymbol{\Theta},\boldsymbol{\eta},\tau,\boldsymbol{T}) >_q) \\
& -\tfrac{<\tau>_\tau}{2}\textstyle\sum_s q^{opt}(s)||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\mu}_s)||^2 \\
& -\tfrac{<\tau>_\tau}{2}\textstyle\sum_s q^{opt}(s)\boldsymbol{W}_s^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{W}_s :< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}|s} \\
& -\tfrac{<\tau>_\tau}{2}\textstyle\sum_s q^{opt}(s)\boldsymbol{G}_s^T\boldsymbol{G}_s :< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{\boldsymbol{\eta}|s} \\
& +\textstyle\sum_s q^{opt}(s)\log\tfrac{1}{S} && (< \log p_s(s) >_q) \\
& +(a_0-1) < \log\tau >_\tau - b_0 < \tau >_\tau && (< \log p_\tau(\tau) >_\tau) \\
& +\textstyle\sum_s q^{opt}(s)(\tfrac{1}{2}\log|\boldsymbol{\Lambda}_{0,s}| - \tfrac{1}{2}\boldsymbol{\Lambda}_0 :< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}|s}) && (< \log p_\Theta(\boldsymbol{\Theta}|s) >_q) \\
& +\textstyle\sum_s q^{opt}(s)(\tfrac{d_\Psi}{2}\log\lambda_{0,\eta,s} - \tfrac{\lambda_{0,\eta,s}}{2}\boldsymbol{I} :< \boldsymbol{\eta}\boldsymbol{\eta}^T >_{\boldsymbol{\eta}|s}) && (< \log p_\eta(\boldsymbol{\eta}|s) >_q) \\
& -\textstyle\sum_s q^{opt}(s)\tfrac{1}{2}\log|\boldsymbol{\Lambda}_s| && (- < \log q^{opt}(\boldsymbol{\Theta}|s) >_q) \\
& -\textstyle\sum_s q^{opt}(s)\tfrac{d_\Psi}{2}\log\lambda_{\eta,s} && (- < \log q^{opt}(\boldsymbol{\eta}|s) >_q) \\
& -\textstyle\sum_s q^{opt}(s)\log q^{opt}(s) && (- < \log q^{opt}(s) >_s) \\
& -(a-1) < \log\tau >_\tau + b < \tau >_\tau + \log Z(a,b), && (- < \log q^{opt}(\tau) >_\tau)
\end{aligned}
\tag{B.1}
$$

where $Z(a,b) = \frac{\Gamma(a)}{b^a}$ is the normalization constant of a $Gamma$ distribution with parameters a, b.

Certain terms become constants and can be neglected. By reformulating, we can derive (see also Equation (4.27), Equation (4.25), Equation (4.26)):

$$
((a_0-1)+\frac{d_y}{2}-(a-1)) < \log\tau >_\tau = (a_0+\frac{d_y}{2}-a) < \log\tau >_\tau = 0, \tag{B.2}
$$

$$
\begin{aligned}
& -\tfrac{1}{2}\textstyle\sum_s q^{opt}(s)((< \tau >_\tau \boldsymbol{W}_s^T\boldsymbol{G}_s^T\boldsymbol{G}_s\boldsymbol{W}_s + \boldsymbol{\Lambda}_0) :< \boldsymbol{\Theta}\boldsymbol{\Theta}^T >_{\boldsymbol{\Theta}|s}) \\
& = -\tfrac{1}{2}\textstyle\sum_s q^{opt}(s)\boldsymbol{\Lambda}_s : \boldsymbol{\Lambda}_s^{-1} \\
& = -\tfrac{d_\Psi}{2},
\end{aligned}
\tag{B.3}
$$

$$
\begin{aligned}
&-\tfrac{1}{2}\textstyle\sum_s q^{opt}(s)(<\tau>_\tau \boldsymbol{G}_s^T\boldsymbol{G}_s :<\boldsymbol{\eta}\boldsymbol{\eta}^T>_{\boldsymbol{\eta}|s} +\lambda_{0,\eta,s}\boldsymbol{I} :<\boldsymbol{\eta}\boldsymbol{\eta}^T>_{\boldsymbol{\eta}|s})\\
&=-\tfrac{1}{2}\textstyle\sum_s q^{opt}(s)\lambda_{\eta,s}\lambda_{\eta,s}^{-1}d_\Psi,\\
&=-\tfrac{d_\Psi}{2}.
\end{aligned}
\tag{B.4}
$$

$$
\begin{aligned}
&-b_0<\tau>_\tau +b<\tau>_\tau +\log Z(a,b)\\
&=-b_0<\tau>_\tau +b\tfrac{a}{b}+\log(\tfrac{\Gamma(a)}{b^a})\\
&=-b_0<\tau>_\tau +a+\log(\Gamma(a))-a\log(b)\\
&=-b_0<\tau>_\tau +a+\log(\Gamma(a))-a\log(\tfrac{a}{<\tau>_\tau})\\
&=-b_0<\tau>_\tau +a+\log(\Gamma(a))-a\log(a)+a\log(<\tau>_\tau)\\
&\propto a\log(<\tau>_\tau),
\end{aligned}
\tag{B.5}
$$

as a from Equation (4.27) is constant, $b_0=0$ and $<\tau>_\tau=\frac{a}{b}$.

Therefore, Equation (B.1) becomes (neglecting constant terms and including Equation (4.28)):

$$
\begin{aligned}
\hat{\mathcal{F}}(q^{opt}(\boldsymbol{\Theta},\boldsymbol{\eta},\tau,s),\boldsymbol{T}) = \;& \textstyle\sum_s q^{opt}(s)[-\frac{<\tau>_\tau}{2}||\hat{\boldsymbol{y}}-\boldsymbol{y}(\boldsymbol{\mu}_s)||^2\\
&+\tfrac{1}{2}\log\tfrac{|\boldsymbol{\Lambda}_{0,s}|}{|\boldsymbol{\Lambda}_s|}+\tfrac{d_\Psi}{2}\log\tfrac{\lambda_{0,\eta,s}}{\lambda_{\eta,s}}-\log q^{opt}(s)]\\
&+a\log(<\tau>_\tau).
\end{aligned}
\tag{B.6}
$$

Appendix C

Determination of required number of basis vectors - Adaptive learning for MoG

An important question is how many basis vectors in $\boldsymbol{W}_j \in \mathbb{R}^{d_\Psi \times d_\Theta}$ should be considered for a mixture component j. We use an information-theoretic criterion Section 3.2.5 that measures the information gain of the approximated posterior to the prior beliefs. Specifically, if $p_{d_\Theta}(\boldsymbol{\Theta}|s)$ (Equation (4.10)) and $q_{d_\Theta}(\boldsymbol{\Theta}|s)$ (Equation (4.24)) denote the $d_\Theta-$dimensional prior and posterior for a given $s = j$, we define the quantity $I(d_\Theta, s)$ as follows:

$$I(d_\Theta, s) = \frac{KL(p_{d_\Theta}(\boldsymbol{\Theta}|s)||q_{d_\Theta}(\boldsymbol{\Theta}|s)) - KL(p_{d_\Theta-1}(\boldsymbol{\Theta}|s)||q_{d_\Theta-1}(\boldsymbol{\Theta}|s))}{KL(p_{d_\Theta}(\boldsymbol{\Theta}|s)||q_{d_\Theta}(\boldsymbol{\Theta}|s))}, \tag{C.1}$$

which measures the (relative) information gain from $d_\Theta - 1$ to d_Θ reduced coordinates. When the information gain falls below a threshold I_{max}, we assume that the information gain is marginal and the addition of reduced coordinates can be terminated. For all mixture components we consider the same d_Θ, chosen from the mixture component that requires the largest d_Θ. Therefore, d_Θ is determined when the information gain with respect to all mixture components falls below the threshold I_{max}, (in our examples we use $I_{max} = 1\%$):

$$max(I(d_\Theta, s = 1), I(d_\Theta, s = 2), ..., I(d_\Theta, s = S)) \leq I_{max}. \tag{C.2}$$

The KL divergence between the two Gaussians, $p_{d_\Theta}(\boldsymbol{\Theta}|s) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_{0,s}^{-1})$ and $q_{d_\Theta}(\boldsymbol{\Theta}|s) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_s^{-1})$, where $\boldsymbol{\Lambda}_{0,s}^{-1}$ and $\boldsymbol{\Lambda}_s^{-1}$ are diagonal, Equation (4.42), is given by:

$$KL(p_{d_\Theta}(\boldsymbol{\Theta}|s)||q_{d_\Theta}(\boldsymbol{\Theta}|s)) = \frac{1}{2}\sum_{i=1}^{d_\Theta}(-\log(\frac{\lambda_{s,i}}{\lambda_{0,s,i}}) + \frac{\lambda_{s,i}}{\lambda_{0,s,i}} - 1), \tag{C.3}$$

and (Equation (C.1)) becomes:

$$I(d_\Theta, s) = \frac{\sum_{i=1}^{d_\Theta}(-\log(\frac{\lambda_{s,i}}{\lambda_{0,s,i}}) + \frac{\lambda_{s,i}}{\lambda_{0,s,i}} - 1) - \sum_{i=1}^{d_\Theta - 1}(-\log(\frac{\lambda_{s,i}}{\lambda_{0,s,i}}) + \frac{\lambda_{s,i}}{\lambda_{0,s,i}} - 1)}{\sum_{i=1}^{d_\Theta}(-\log(\frac{\lambda_{s,i}}{\lambda_{0,s,i}}) + \frac{\lambda_{s,i}}{\lambda_{0,s,i}} - 1)}. \tag{C.4}$$

Naturally, one could consider different values of d_Θ for each mixture component which could lead to additional savings.

Appendix D

Computational cost

	formulary /costs	costs
μ-update $\mathcal{F}_\mu(\boldsymbol{\mu})$:	$\Delta\boldsymbol{y}^T\Delta\boldsymbol{y}$ $\quad p_\mu(\boldsymbol{\mu})$ $2d_y$	-
Deriving $\boldsymbol{\mu}$:	$\boldsymbol{G}^T\boldsymbol{G}$ $\quad ()^{-1}()$ $\quad \boldsymbol{G}^T\Delta\boldsymbol{y}$ $\quad \boldsymbol{G}^T\boldsymbol{G}\boldsymbol{\mu}$ $2d_y d_\Psi^2$ $\quad 2/3d_\Psi^3$ $\quad 2d_y d_\Psi$ $\quad 2d_\Psi^2$	$2d_y d_\Psi^2 + 2/3d_\Psi^3$
Forward call with derv.:	Xd_y^3	Xd_y^3
W-update $\mathcal{F}_W(\boldsymbol{W})$:	$\boldsymbol{W}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W} : \boldsymbol{\Lambda}^{-1}$ $2d_\Psi^2 d_\Theta + 2d_\Psi d_\Theta + 2d_\Theta$	-
$\partial\mathcal{F}_W/\partial\boldsymbol{W}$:	$\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{W} : \boldsymbol{\Lambda}^{-1}$ -	-
Cayley-update:	$4d_\Psi d_\Theta^2 + \mathcal{O}(d_\Theta^3)$	-
q-update	Neglectible	-

Table D.1: It is $d_\Theta \ll d_\Psi$. The major costs within each sup-iteration is listed in the right column, if the order of dimensions is of power three. We do not consider contributions from d_Θ as it is much smaller than the other dimensions. Costs are only listed if they arise within each sub-iteration. The cost for $\partial\mathcal{F}_W/\partial\boldsymbol{W}$ occur either just once and does not need to be recalculated within the next update, or the product of the matrices have already been derived when calculating $\mathcal{F}_W(\boldsymbol{W})$. The computational cost for the forward call including the derivation of the derivatives depend on the application (e.g., if a linear or nonlinear system (which then may be solved iteratively) is included). Therefore, the cost are indicated with an estimate of Xd_y^3, where X is a factor which depends on the specific system.

Appendix E

Numerical implementation

The numerical implementation of the main solver has been build in C++, based on the 'boost'-library [230].

The finite element simulations for the forward simulations $\boldsymbol{y}(\boldsymbol{\Psi})$ and the derivatives $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}$ are conducted with the freely available finite element software for biomechanics 'FEBio' [231] (release version 1.6.1). The C++ software has been developed by the Musculoskeletal Research Laboratory at the University of Utah (USA) and is a nonlinear finite element solver, specified for biomechanics and biophysics. Different material models are available, such as in solid mechanics: Neo-Hookean, Mooney-Rivlin, Odgen or the Veronda-Westmann model. Graduate students of the Continuum Mechanics group at the TUM extended FEBio with regard to derivatives $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\Psi}}$ for the purpose of this thesis accordingly.

The FEBio code is used as a black box which interacts iteratively with the main solver. The strict separation between the new developed and implemented methods and the applications underlies the fast possible adaption for other requests.

Appendix F

Verification with Gibbs sampling

E-step - Gibbs Sampling

To derive expectations with regard to the posterior we carry out Gibbs sampling with respect to each of the components of $\boldsymbol{\Upsilon}$, i.e., $\boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\sigma}, \boldsymbol{H}$. This requires conditional distributions of the parameters, using the other conditionally sampled parameters of $\boldsymbol{\Upsilon}^{(i)}$, where the superscript i denotes the specific sample, $i = 1 : N$:

- **For** $p(\boldsymbol{y} \mid \boldsymbol{\sigma}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Lambda}_y^{-1})$

$$\boldsymbol{\Lambda}_y = \sum_{e=1}^{d_{FE}} v_e \Psi_e^{(i)} \Psi_e^{(i)} \boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}} \boldsymbol{B}_{y,e} \boldsymbol{L}_{y,e} + \tau \boldsymbol{I}_{d_y}, \tag{F.1}$$

$$\boldsymbol{\mu}_y = \boldsymbol{\Lambda}_y^{-1} [\sum_{e=1}^{d_{FE}} v_e \Psi_e^{(i)} \boldsymbol{L}_{y,e}^T \boldsymbol{B}_{y,e}^T \tilde{\boldsymbol{D}}^T (\boldsymbol{\sigma}_e^{(i)} - \Psi_e^{(i)} \tilde{\boldsymbol{D}} \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e}) + \tau \hat{\boldsymbol{y}}]. \tag{F.2}$$

  $\boldsymbol{y}^{(i)}$ can then be sampled from $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Lambda}_y^{-1})$.

- **For** $p(\boldsymbol{\Psi} \mid \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{y}, \boldsymbol{H}) = \mathcal{N}(\bar{\boldsymbol{\mu}}_\Psi, \bar{\boldsymbol{\Lambda}}_\Psi^{-1})$
  For the material parameters the conditional probability distributions follow with:

$$\Lambda_{\Psi,e} = v_e (\boldsymbol{B}_{y,e} \boldsymbol{y}_e^{(i)} + \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e})^T \tilde{\boldsymbol{D}}_e^T \tilde{\boldsymbol{D}}_e (\boldsymbol{B}_{y,e} \boldsymbol{y}_e^{(i)} + \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e}), \tag{F.3}$$

  and

$$\Lambda_{\Psi,e} \mu_{\Psi,e} = v_e (\boldsymbol{B}_{y,e} \boldsymbol{y}_e^{(i)} + \boldsymbol{B}_{b,e} \boldsymbol{u}_{b,e}) \tilde{\boldsymbol{D}}_e^T \boldsymbol{\sigma}_e^{(i)}. \tag{F.4}$$

  Including a smoothing prior, compare Equation (5.20), the variance and mean result in:

$$\bar{\boldsymbol{\Lambda}}_\Psi = \boldsymbol{\Lambda}_\Psi + (\boldsymbol{L}_\Psi^T \boldsymbol{H}^{(i)} \boldsymbol{L}_\Psi), \tag{F.5}$$

  and

$$\bar{\boldsymbol{\mu}}_\Psi = \bar{\boldsymbol{\Lambda}}_\Psi^{-1} (\boldsymbol{\Lambda}_\Psi \boldsymbol{\mu}_\Psi). \tag{F.6}$$

  $\boldsymbol{\Psi}^{(i)}$ is sampled from $\mathcal{N}(\bar{\boldsymbol{\mu}}_\Psi, \bar{\boldsymbol{\Lambda}}_\Psi^{-1})$.

- **For** $p(\boldsymbol{H} \mid \boldsymbol{\Psi}) = \prod_{l=1}^{d_L} Gamma(a_{h,l}, b_{h,l})$
  When including a smoothing prior the hyperparameters h_l, conditional on the other parameters, follow a Gamma distribution $Gamma(a_{h,l}, b_{h,l})$ with

$$a_{h,l} = a_{h,0} + \frac{1}{2}, \tag{F.7}$$

$$b_{h,l} = b_{h,0} + \frac{1}{2}(\Psi_{l,1}^{(i)} - \Psi_{l,2}^{(i)})^2. \tag{F.8}$$

  $h_l^{(i)}$ is sampled from $Gamma(a_{h,l}, b_{h,l})$ for all l which results in $\boldsymbol{H}^{(i)}$.

- **For** $p(\boldsymbol{\sigma} \mid \boldsymbol{V}, \boldsymbol{\Psi}, \boldsymbol{y}) = \mathcal{N}(\bar{\boldsymbol{\mu}}_\sigma, \bar{\boldsymbol{\Lambda}}_\sigma^{-1})$

$$\boldsymbol{\Lambda}_{\sigma,e} = v_e \boldsymbol{I}_{d_{S_e}}, \tag{F.9}$$

$$\boldsymbol{\Lambda}_{\sigma,e}\boldsymbol{\mu}_{\sigma,e} = v_e \Psi_e^{(i)} \tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_e^{(i)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e}). \tag{F.10}$$

  Including the prior information of the equilibrium constrain the updated conditional distributions follow with:

$$\bar{\boldsymbol{\Lambda}}_\sigma = k\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T + \boldsymbol{\Lambda}_\sigma, \tag{F.11}$$

$$\bar{\boldsymbol{\mu}}_\sigma = \bar{\boldsymbol{\Lambda}}_\sigma^{-1}(k\hat{\boldsymbol{B}}\boldsymbol{f} + \boldsymbol{\Lambda}_\sigma\boldsymbol{\mu}_\sigma). \tag{F.12}$$

  Then $\boldsymbol{\sigma}^{(i)}$ is sampled from $\mathcal{N}(\bar{\boldsymbol{\mu}}_\sigma, \bar{\boldsymbol{\Lambda}}_\sigma^{-1})$.

M-step - Deriving v

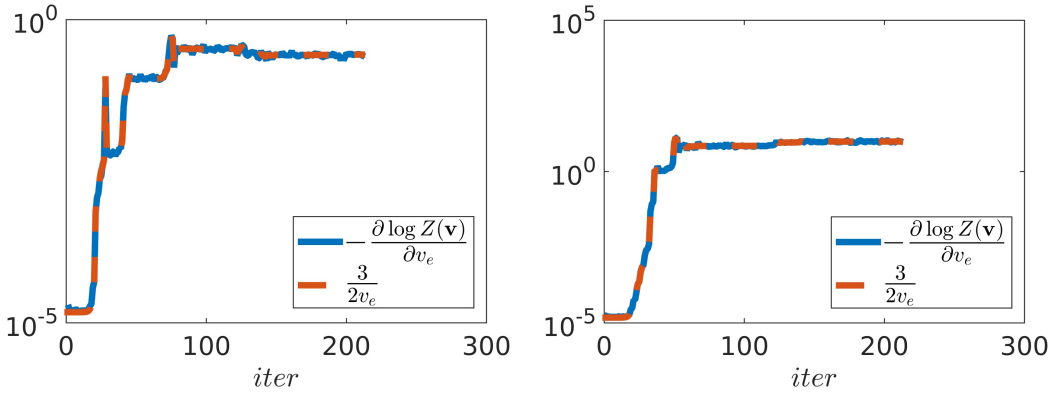
The update equations follow Section 5.2.1, only for Equation (5.51) the required expectation is derived by: where Ξ_e, the expectation, can be derived from the Gibbs samples by

$$\begin{aligned} \Xi_e &= < ||\boldsymbol{\sigma}_e - \Psi_e \tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_e + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2 >_q \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} ||\boldsymbol{\sigma}_e^{(i)} - \Psi_e^{(i)} \tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{y,e}\boldsymbol{y}_e^{(i)} + \boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e})||^2. \end{aligned} \tag{F.13}$$

Appendix G

Approximation of normalization constant

In Figure G.1, for two exemplary finite elements the evolution of the derivaties $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ (Equation (5.66)) over the number of iterations is shown. One of the selected elements, $e = 85$, has no and the other element has model error, $e = 549$. In the figure, it is exemplarily shown that the evolution of $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ coincides with $\frac{3}{2v_e}$.



Element 85 Element 549

Figure G.1: Evolution of $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ in comparison with $\frac{3}{2v_e}$, for the elements 85 (without model error) and 549 (with model error).

In Figure G.2, in addition to the expected values, we show $-\frac{1}{2}||\boldsymbol{\sigma}_{z,e}^{(ii)} - \Psi_{z,e}^{(ii)}\tilde{\boldsymbol{D}}_e(\boldsymbol{B}_{b,e}\boldsymbol{u}_{b,e} + \boldsymbol{B}_{y,e}\boldsymbol{y}_{z,e}^{(ii)})||^2$ over the number of samples. Furthermore, the ergodic mean of the samples (in red), the derived expectation (in blue) as well as the estimated normalization constant of $\frac{3}{2v_e}$ is shown.

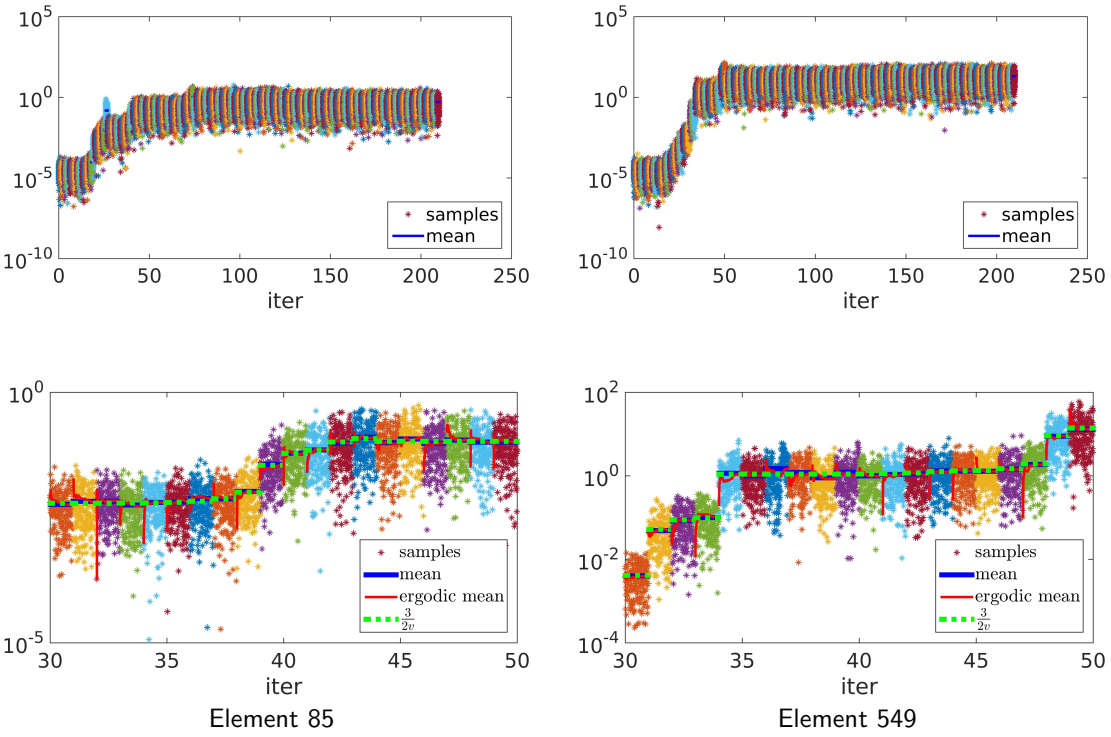


Figure G.2: Development of $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ over the number of iteration for the elements 85 and 549. This is compared to $\frac{3}{2v_e}$, which coincides.

We also run examples where $\boldsymbol{v}$ is not derived as discussed but with the assumption of $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e} = \frac{3}{2v_e}$. In the resulting outcome, we could not observe any differences. However, since we were not able to prove this relation mathematically, we recommend to start with $\frac{3}{2v_e}$ at the beginning. Then, at the end of the simulations Gibbs sampling should be employed to ensure convergence to the correct result. This has, besides the advantage that Gibbs-sampling for $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e}$ is not required, also the advantage that under the assumption of $-\frac{\partial \log Z(\boldsymbol{v})}{\partial v_e} = \frac{3}{2v_e}$ the perfect step-size for updating v_e is available. This accelerates in addition the overall simulation.

Bibliography

[1] Arthur Conan Doyle. A Study in Scarlet. 1887. *London: Ward*, 1888.

[2] George E. Backus and J. F. Gilbert. Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal International*, 13(1-3):247–276, 1967.

[3] Tan Bui-Thanh, Omar Ghattas, James Martin, and Georg Stadler. A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013.

[4] Jean-Luc Starck. Sparsity and inverse problems in astrophysics. In *Journal of Physics: Conference Series*, volume 699, page 012010. IOP Publishing, 2016.

[5] Georgios E. Stavroulakis. *Inverse and crack identification problems in engineering mechanics*, volume 46. Springer Science & Business Media, 2013.

[6] Iain S. Weir. Fully Bayesian reconstructions from single-photon emission computed tomography data. *Journal of the American Statistical Association*, 92(437):49–60, 1997.

[7] Jingbo Wang and Nicholas Zabaras. A Bayesian inference approach to the inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, 47(17):3927–3941, 2004.

[8] Jingbo Wang and Nicholas Zabaras. Hierarchical Bayesian models for inverse problems in heat conduction. *Inverse Problems*, 21(1):183, 2005.

[9] Paul Dostert, Yalchin Efendiev, Thomas Y. Hou, and Wuan Luo. Coarse-gradient Langevin algorithms for dynamic data integration and uncertainty quantification. *Journal of computational physics*, 217(1):123–142, 2006.

[10] Lorenz Biegler, George Biros, Omar Ghattas, Matthias Heinkenschloss, David Keyes, Bani Mallick, Luis Tenorio, Bart van Bloemen Waanders, Karen Willcox, and Youssef Marzouk. *Large-scale inverse problems and quantification of uncertainty*, volume 712. John Wiley & Sons, 2011.

[11] Assad A. Oberai, Nachiket H. Gokhale, Sevan Goenezen, Paul E. Barbone, Timothy J. Hall, Amy M. Sommer, and Jingfeng Jiang. Linear and nonlinear elasticity imaging of soft tissue in vivo: demonstration of feasibility. *Physics in medicine and biology*, 54(5):1191, 2009.

[12] Nathalie Ganne-Carri, Marianne Ziol, Victor de Ledinghen, Catherine Douvin, Patrick Marcellin, Laurent Castera, Daniel Dhumeaux, Jean-Claude Trinchet, and Michel Beaugrand. Accuracy of liver stiffness measurement for the diagnosis of cirrhosis in patients with chronic liver diseases. *Hepatology*, 44(6):1511–1517, 2006.

[13] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, and others. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[14] Erwin Kuntz and Hans-Dieter Kuntz. *Hepatology, Principles and practice: history, morphology, biochemistry, diagnostics, clinic, therapy*. Springer Science & Business Media, 2006.

[15] Parris Wellman, Robert D. Howe, Edward Dalton, and Kenneth A. Kern. Breast tissue stiffness in compression is correlated to histological diagnosis. *Harvard BioRobotics Laboratory Technical Report*, 1999.

[16] Thomas A. Krouskop, Thomas M. Wheeler, Faouzi Kallel, Brian S. Garra, and Timothy Hall. Elastic moduli of breast and prostate tissues under compression. *Ultrasonic imaging*, 20(4):260–274, 1998.

[17] Cdric Schmitt, Gilles Soulez, Roch L. Maurice, Marie-France Giroux, and Guy Cloutier. Noninvasive vascular elastography: toward a complementary characterization tool of atherosclerosis in carotid arteries. *Ultrasound in medicine & biology*, 33(12):1841–1858, 2007.

[18] Jacques Ohayon, Grard Finet, Simon Le Floch, Guy Cloutier, Ahmed M. Gharib, Julie Heroux, and Roderic I. Pettigrew. Biomechanics of atherosclerotic coronary plaque: site, stability and in vivo elasticity modeling. *Annals of biomedical engineering*, 42(2):269–279, 2014.

[19] Spencer W. Shore, Paul E. Barbone, Assad A. Oberai, and Elise F. Morgan. Transversely isotropic elasticity imaging of cancellous bone. *Journal of biomechanical engineering*, 133(6):061002, 2011.

[20] Wen-Chun Yeh, Pai-Chi Li, Yung-Ming Jeng, Hey-Chi Hsu, Po-Ling Kuo, Meng-Lin Li, Pei-Ming Yang, and Po Huang Lee. Elastic modulus measurements of human liver and correlation with pathology. *Ultrasound in medicine & biology*, 28(4):467–474, 2002.

[21] World Health Organization. *World Health Statistics 2014*. WHO Press, Switzerland, 2014.

[22] National Cancer Institute,. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), April 2016.

[23] A. Goddi, M. Bonardi, and S. Alessi. Breast elastography: A literature review. *Journal of Ultrasound*, 15(3):192–198, June 2012.

[24] J. Ophir, I. Cespedes, Hm Ponnekanti, Y. Yazdi, and X. Li. Elastography: a quantitative method for imaging the elasticity of biological tissues. *Ultrasonic imaging*, 13(2):111–134, 1991.

[25] Brian S. Garra, E. Ignacio Cespedes, J. Ophir, Stephen R. Spratt, Rebecca A. Zuurbier, Colette M. Magnant, and Marie F. Pennanen. Elastography of breast lesions: initial clinical results. *Radiology*, 202(1):79–86, 1997.

[26] Jeffrey C. Bamber, Paul E. Barbone, David O. COSGROVE, Marvin M. DOYELY, Frank G. FUECHSEL, Paul M. MEANEY, Naomi R. MILLER, Tsuyoshi SHIINA, Francois TRANQUART, and others. Progress in freehand elastography of the breast. *IEICE TRANSACTIONS on Information and Systems*, 85(1):5–14, 2002.

[27] A. Thomas, T. Fischer, H. Frey, R. Ohlinger, S. Grunwald, J.-U. Blohmer, K.-J. Winzer, S. Weber, G. Kristiansen, B. Ebert, and others. Real-time elastographyan advanced method of ultrasound: first results in 108 patients with breast lesions. *Ultrasound in obstetrics & gynecology*, 28(3):335–340, 2006.

[28] Kevin J. Parker, M. M. Doyley, and D. J. Rubens. Imaging the elastic properties of tissue: the 20 year perspective. *Physics in medicine and biology*, 56(1):R1, 2011.

[29] M. M. Doyley. Model-based elastography: a survey of approaches to the inverse elasticity problem. *Physics in medicine and biology*, 57(3):R35, 2012.

[30] R. Muthupillai, D. J. Lomas, P. J. Rossman, J. F. Greenleaf, A. Manduca, and R. L. Ehman. Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. *Science*, 269(5232):1854–1857, 1995.

[31] Ahmad S. Khalil, Raymond C. Chan, Alexandra H. Chau, Brett E. Bouma, and Mohammad R. Kaazempur Mofrad. Tissue elasticity estimation with optical coherence elastography: toward mechanical characterization of in vivo soft tissue. *Annals of biomedical engineering*, 33(11):1631–1639, 2005.

[32] J.-L. Gennisson, Thomas Deffieux, Mathias Fink, and Michal Tanter. Ultrasound elastography: principles and techniques. *Diagnostic and interventional imaging*, 94(5):487–495, 2013.

[33] Xian Huang, Le-Hang Guo, Hui-Xiong Xu, Xue-Hao Gong, Bo-Ji Liu, Jun-Mei Xu, Yi-Feng Zhang, Xiao-Long Li, Dan-Dan Li, Shen Qu, and others. Acoustic radiation force impulse induced strain elastography and point shear wave elastography for evaluation of thyroid nodules. *International journal of clinical and experimental medicine*, 8(7):10956, 2015.

[34] Armen Sarvazyan, Timothy J Hall, Matthew W Urban, Mostafa Fatemi, Salavat R Aglyamov, and Brian S Garra. An overview of elastography-an emerging branch of medical imaging. *Current medical imaging reviews*, 7(4):255–282, 2011.

[35] Joel Edward Lindop. *2D and 3D elasticity imaging using freehand ultrasound*. PhD thesis, University of Cambridge, 2008.

[36] Paul E. Barbone, Carlos E. Rivas, Isaac Harari, Uri Albocher, Assad A. Oberai, and Yixiao Zhang. Adjoint-weighted variational formulation for the direct solution of inverse problems of general linear elasticity with full interior data. *International journal for numerical methods in engineering*, 81(13):1713–1736, 2010.

[37] Peter R. Hoskins, Kevin Martin, and Abigail Thrush. *Diagnostic ultrasound: physics and equipment*. Cambridge University Press, 2010.

[38] Assad A. Oberai, Nachiket H. Gokhale, Marvin M. Doyley, and Jeffrey C. Bamber. Evaluation of the adjoint equation based algorithm for elasticity imaging. *Physics in Medicine and Biology*, 49(13):2955, 2004.

[39] Marvin M. Doyley, Seshadri Srinivasan, Eugene Dimidenko, Nirmal Soni, and Jonathan Ophir. Enhancing the performance of model-based elastography by incorporating additional a priori information in the modulus image reconstruction process. *Physics in medicine and biology*, 51(1):95, 2006.

[40] Alexander Arnold, Stefan Reichling, Otto T. Bruhns, and Jrn Mosler. Efficient computation of the elastography inverse problem by combining variational mesh adaption and a clustering technique. *Physics in Medicine and Biology*, 55(7):2035, 2010.

[41] Lorraine G. Olson and Robert D. Throne. Numerical simulation of an inverse method for tumour size and location estimation. *Inverse Problems in Science and Engineering*, 18(6):813–834, 2010.

[42] DS SCHNUR and N. ZABARAS. An inverse method for determining elastic material properties and a material interface. *International journal for numerical methods in engineering*, 33(10):2039–2057, 1992.

[43] Nachiket H. Gokhale, Paul E. Barbone, and Assad A. Oberai. Solution of the nonlinear elasticity imaging inverse problem: the compressible case. *Inverse Problems*, 24(4):045010, August 2008.

[44] Sevan Goenezen, Jean-Francois Dord, Zac Sink, Paul E. Barbone, Jingfeng Jiang, Timothy J. Hall, and Assad A. Oberai. Linear and nonlinear elastic modulus imaging: an application to breast cancer diagnosis. *IEEE transactions on medical imaging*, 31(8):1628–1637, 2012.

[45] Curtis R. Vogel. *Computational methods for inverse problems*, volume 23. Siam, 2002.

[46] Marc C. Kennedy and Anthony O'Hagan. Bayesian Calibration of Computer Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(3):425–464, January 2001.

[47] Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 2012.

[48] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2003.

[49] Peter J. Green, Krzysztof Latuszyski, Marcelo Pereyra, and Christian P. Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862, 2015.

[50] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[51] Gareth O. Roberts, Jeffrey S. Rosenthal, and others. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[52] Jonathan C. Mattingly, Natesh S. Pillai, Andrew M. Stuart, and others. Diffusion limits of the random walk Metropolis algorithm in high dimensions. *The Annals of Applied Probability*, 22(3):881–930, 2012.

[53] Natesh S. Pillai, Andrew M. Stuart, Alexandre H. Thiry, and others. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012.

[54] Christopher M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[55] Herbert KH Lee, Dave M. Higdon, Zhuoxin Bi, Marco AR Ferreira, and Mike West. Markov random field models for high-dimensional parameters in simulations of fluid flow in porous media. *Technometrics*, 44(3):230–241, 2002.

[56] Christopher H. Holloman, Herbert KH Lee, and Dave M. Higdon. Multi-resolution Genetic Algorithms and Markov Chain Monte Carlo. Technical report, Technical report, Duke University, 2002.

[57] Nicolas Chopin, Tony Lelivre, and Gabriel Stoltz. Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Statistics and Computing*, 22(4):897–916, 2012.

[58] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[59] Phaedon-Stelios Koutsourelakis. A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters. *Journal of computational physics*, 228(17):6184–6211, 2009.

[60] Pierre Del Moral, Arnaud Doucet, Ajay Jasra, and others. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278, 2012.

[61] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[62] James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[63] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal on Scientific Computing*, 36(4):A1525–A1555, 2014.

[64] Tan Bui-Thanh and Mark Girolami. Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo. *Inverse Problems*, 30(11):114014, 2014.

[65] David Galbally, K. Fidkowski, K. Willcox, and O. Ghattas. Nonlinear model reduction for uncertainty quantification in large-scale inverse problems. 2009.

[66] Youssef M. Marzouk, Habib N. Najm, and Larry A. Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, June 2007.

[67] Tan Bui-Thanh, Karen Willcox, and Omar Ghattas. Model reduction for large-scale systems with high-dimensional parametric input space. *SIAM Journal on Scientific Computing*, 30(6):3270–3288, 2008.

[68] Bojana V. Rosi, Alexander Litvinenko, Oliver Pajonk, and Hermann G. Matthies. Sampling-free linear Bayesian update of polynomial chaos representations. *Journal of Computational Physics*, 231(17):5761–5787, 2012.

[69] I. Bilionis and N. Zabaras. Solution of inverse problems with limited forward solver evaluations: a Bayesian perspective. *Inverse Problems*, 30(1):015004, 2014.

[70] P. Chen and Ch Schwab. Sparse-grid, reduced-basis bayesian inversion: Nonaffine-parametric nonlinear equations. In *ETH Zurich, Seminar for Applied Mathematics, Report*, volume 21, 2015.

[71] Shiwei Lan, Tan Bui-Thanh, Mike Christie, and Mark Girolami. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian Inverse Problems. *Journal of Computational Physics*, 308:81–101, 2016.

[72] Jingbo Wang and Nicholas Zabaras. Using Bayesian statistics in the estimation of heat source in radiation. *International Journal of Heat and Mass Transfer*, 48(1):15–29, 2005.

[73] Tiangang Cui, Youssef Marzouk, and Karen Willcox. Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction. *Journal of Computational Physics*, 315:363–387, 2016.

[74] Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C. Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, page 3. IEEE Computer Society Press, 2012.

[75] Tiangang Cui, James Martin, Youssef M. Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014.

[76] Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

[77] Tiangang Cui, Kody JH Law, and Youssef M. Marzouk. Dimension-independent likelihood-informed MCMC. *Journal of Computational Physics*, 304:109–137, 2016.

[78] H. Pearl Flath, Lucas C. Wilcox, Volkan Akelik, Judith Hill, Bart van Bloemen Waanders, and Omar Ghattas. Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. *SIAM Journal on Scientific Computing*, 33(1):407–432, 2011.

[79] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. University of London, 2003.

[80] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[81] Hagai Attias. A variational Bayesian framework for graphical models. *Advances in neural information processing systems*, 12(1-2):209–215, 2000.

[82] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[83] Michael Chappell, Adrian R. Groves, Brandon Whitcher, Mark W. Woolrich, and others. Variational Bayesian inference for a nonlinear forward model. *Signal Processing, IEEE Transactions on*, 57(1):223–236, 2009.

[84] Bangti Jin and Jun Zou. Hierarchical Bayesian inference for ill-posed problems via variational method. *Journal of Computational Physics*, 229(19):7317–7343, 2010.

[85] T. M. Cover and J. A. Thomas. *Elements of information theory*. 1991. INSPEC:4044734.

[86] Tarek A. El Moselhy and Youssef M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, October 2012.

[87] Jeff Gill and George Casella. Dynamic tempered transitions for exploring multimodal posterior distributions. *Political Analysis*, 12(4):425–443, 2004.

[88] Weixuan Li and Guang Lin. An adaptive importance sampling algorithm for Bayesian inversion with multimodal distributions. *Journal of Computational Physics*, 294:173–190, 2015.

[89] Farhan Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, 2008.

[90] Bin Liu. Adaptive annealed importance sampling for multimodal posterior exploration and model selection with application to extrasolar planet detection. *The Astrophysical Journal Supplement Series*, 213(1):14, 2014.

[91] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(13):19–41, January 2000.

[92] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[93] R. A. Choudrey and S. J. Roberts. Variational mixture of Bayesian independent component analyzers. *Neural Computation*, 15(1):213–252, January 2003.

[94] Mikael Kuusela, Tapani Raiko, Antti Honkela, and Juha Karhunen. A gradient-based algorithm competitive with variational Bayesian EM for mixture of Gaussians. In *2009 International Joint Conference on Neural Networks*, pages 1688–1695. IEEE, 2009.

[95] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[96] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[97] Gideon Schwarz and others. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[98] Jenny Brynjarsdottir and Anthony O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30(11):114007, 2014.

[99] Dave Higdon, Charles Nakhleh, James Gattiker, and Brian Williams. A Bayesian calibration approach to the thermal problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29):2431–2441, 2008.

[100] L. Mark Berliner, Kenneth Jezek, Noel Cressie, Yongku Kim, C. Q. Lam, and Cornelis J. van der Veen. Modeling dynamic controls on ice streams: a Bayesian statistical approach. *Journal of Glaciology*, 54(187):705–714, 2008.

[101] Michael Emory, Johan Larsson, and Gianluca Iaccarino. Modeling of structural uncertainties in Reynolds-averaged Navier-Stokes closures. *Physics of Fluids (1994-present)*, 25(11):110822, 2013.

[102] Jin-Long Wu, Jian-Xun Wang, and Heng Xiao. A Bayesian CalibrationPrediction Method for Reducing Model-Form Uncertainties with Application in RANS Simulations. *Flow, Turbulence and Combustion*, pages 1–26, 2015.

[103] H. Xiao, J.-L. Wu, J.-X. Wang, R. Sun, and C. J. Roy. Quantifying and reducing model-form uncertainties in Reynolds-averaged Navier-Stokes simulations: A data-driven, physics-informed Bayesian approach. *Journal of Computational Physics*, 324:115–136, 2016.

[104] Eric Dow and Qiqi Wang. Quantification of structural uncertainties in the k-omega turbulence model. *AIAA Paper*, 1762:2011, 2011.

[105] Eric J. Parish and Karthik Duraisamy. A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics*, 305:758–774, 2016.

[106] Anand Pratap Singh and Karthik Duraisamy. Using field inversion to quantify functional errors in turbulence closures. *Physics of Fluids (1994-present)*, 28(4):045110, 2016.

[107] Todd A. Oliver and Robert D. Moser. Bayesian uncertainty quantification applied to RANS turbulence models. In *Journal of Physics: Conference Series*, volume 318, page 042032. IOP Publishing, 2011.

[108] Phaedon-Stelios Koutsourelakis. A novel Bayesian strategy for the identification of spatially varying material properties and model validation: an application to static elastography. *International Journal for Numerical Methods in Engineering*, 91(3):249–268, 2012.

[109] Panagiotis Tsilifis, Ilias Bilionis, Ioannis Katsounaros, and Nicholas Zabaras. Computationally Efficient Variational Approximations for Bayesian Inverse Problems. *Journal of Verification, Validation and Uncertainty Quantification*, 1(3):031004–031004, July 2016.

[110] Matthew McCormick, Nicholas Rubert, and Tomy Varghese. Bayesian regularization applied to ultrasound strain imaging. *IEEE Transactions on Biomedical Engineering*, 58(6):1612–1620, 2011.

[111] Petter Risholm, Eigil Samset, and William Wells III. Bayesian estimation of deformation and elastic parameters in non-rigid registration. In *International Workshop on Biomedical Image Registration*, pages 104–115. Springer, 2010.

[112] Ivor JA Simpson, Julia A. Schnabel, Adrian R. Groves, Jesper LR Andersson, and Mark W. Woolrich. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59(3):2438–2451, 2012.

[113] I. J. A. Simpson, M. J. Cardoso, M. Modat, D. M. Cash, M. W. Woolrich, J. L. R. Andersson, J. A. Schnabel, and S. Ourselin. Probabilistic non-linear registration with spatially adaptive regularisation. *Medical Image Analysis*, 26(1):203–216, 2015.

[114] Petter Risholm, James Ross, George R. Washko, and William M. Wells. Probabilistic elastography: estimating lung elasticity. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 699–710. Springer, 2011.

[115] Miguel A. Aguilo, Laura Swiler, and Angel Urbina. An overview of inverse material identification within the frameworks of deterministic and stochastic parameter estimation. *International Journal for Uncertainty Quantification*, 3(4), 2013.

[116] John W. Tukey and others. The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[117] Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 11 edition, 2007.

[118] Michael Havbro Faber. *Statistics and Probability Theory: In pursuit of engineering decision support*, volume 18. Springer Science & Business Media, 2012.

[119] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2004.

[120] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[121] Mr Bayes and Mr Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions (1683-1775)*, pages 370–418, 1763.

[122] Pierre Simon marquis de Laplace. *Theorie analytique des probabilites*. V. Courcier, 1820.

[123] R. A. Choudrey. *Variational Methods for Bayesian Independent Component Analysis*. PhD thesis, 2002.

[124] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[125] Ralph C. Smith. *Uncertainty quantification: theory, implementation, and applications*, volume 12. SIAM, 2013.

[126] Pierre Simon Laplace. Memoires de Mathematique et de Physique, Tome Sixieme. Memoir on the probability of the causes of events. (English translation by S. M. Stigler 1986.). *Statistical Science*, 1(3):364–378, 1774.

[127] D. J. C. MacKay. Choice of basis for Laplace approximation. *Machine Learning*, 33(1):77–86, October 1998.

[128] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2001.

[129] Arnaud Doucet, Nando De Freitas, and Neil Gordon. *Sequential Monte Carlo methods in practice*. Springer, 2001.

[130] Z. Ghahramani and M.J. Beal. Graphical models and variational methods. *Advanced Mean Field Methods - Theory and practice. MIT Press*, 2000.

[131] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[132] Christian Robert and George Casella. *Introducing Monte Carlo Methods with R*. Springer Science & Business Media, 2009.

[133] M. Giaquinta and S. Hildebrandt. *Calculus of Variations, vol. I. A Series of Comprehensive Studies in Mathematics, vol. 310*. Springer-Verlag, Berlin, 1996.

[134] Djc Mackay. Probable Networks and Plausible Predictions - a Review of Practical Bayesian Methods for Supervised Neural Networks. *Network-Computation in Neural Systems*, 6(3):469–505, August 1995.

[135] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, volume 89, pages 355–368. 1998.

[136] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les ingalits entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[137] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[138] G. Parisi. *Statistical field theory*. 1988. INSPEC:3450197.

[139] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[140] Gerhard A. Holzapfel. *Nonlinear solid mechanics*, volume 24. Wiley Chichester, 2000.

[141] G. Thomas Mase, Ronald E. Smelser, and George E. Mase. *Continuum mechanics for engineers*. CRC press, 2009.

[142] J. Bonet, A. J. Gil, and R. D. Wood. *Worked Examples in Nonlinear Continuum Mechanics for Finite Element Analysis*. 2012.

[143] Michael W. Gee, Ch Frster, and W. A. Wall. A computational strategy for prestressing patient-specific biomechanical problems under finite deformation. *International Journal for Numerical Methods in Biomedical Engineering*, 26(1):52–72, 2010.

[144] Olgierd Cecil Zienkiewicz, Robert Leroy Taylor, Olgierd Cecil Zienkiewicz, and Robert Lee Taylor. *The finite element method*, volume 3. McGraw-hill London, 1977.

[145] Thomas J. R. Hughes and Ted Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Dover Pubn Inc, Mineola, NY, August 2000.

[146] Sevan Goenezen, Paul Barbone, and Assad A. Oberai. Solution of the nonlinear elasticity imaging inverse problem: The incompressible case. *Computer methods in applied mechanics and engineering*, 200(13):1406–1420, 2011.

[147] E. Weinan. *Principles of multiscale modeling*. Cambridge University Press, 2011.

[148] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical recipes*. Cambridge University Press, Cambridge, 1990.

[149] Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35:67–68, 1999.

[150] Hermann Matthies and Gilbert Strang. The solution of nonlinear finite element equations. *International journal for numerical methods in engineering*, 14(11):1613–1626, 1979.

[151] Klaus-Jrgen Bathe. *Finite element procedures*. Klaus-Jurgen Bathe, 2006.

[152] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. siam, 2005.

[153] Michael B. Giles and Niles A. Pierce. An introduction to the adjoint approach to design. *Flow, turbulence and combustion*, 65(3-4):393–415, 2000.

[154] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. Optimization with PDE Constraints. In *Optimization with Pde Constraints*, volume 23, pages 1–270. 2009.

[155] D. I. Papadimitriou and K. C. Giannakoglou. Direct, adjoint and mixed approaches for the computation of Hessian in airfoil design problems. *International Journal for Numerical Methods in Fluids*, 56(10):1929–1943, 2008.

[156] Kostas Orginos and Andreas Stathopoulos. A solver for multiple right hand sides. *PoS LATTICE*, 42, 2007.

[157] Martin H. Gutknecht and Thomas Schmelzer. The block grade of a block Krylov space. *Linear Algebra and its Applications*, 430(1):174–185, 2009.

[158] George EP Box. Science and Statistics. *Journal of the American Statistical Association*, pages 791–799, 1976.

[159] Isabell M. Franck and P. S. Koutsourelakis. Sparse Variational Bayesian approximations for nonlinear inverse problems: Applications in nonlinear elastography. *Computer Methods in Applied Mechanics and Engineering*, 299:215–244, 2016.

[160] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

[161] Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[162] Matthias W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9(Apr):759–813, 2008.

[163] Ian J. Goodfellow, Aaron Courville, and Yoshua Bengio. Spike-and-slab sparse coding for unsupervised feature discovery. *arXiv preprint arXiv:1201.3382*, 2012.

[164] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

[165] Daniela Calvetti and Lothar Reichel. Tikhonov regularization of large linear problems. *BIT Numerical Mathematics*, 43(2):263–283, 2003.

[166] Johnathan M. Bardsley. Gaussian Markov random field priors for inverse problems. *Inverse Probl. Imaging*, 7(2):397–416, 2013.

[167] C. Schwab and A. M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Problems*, 28(4):045003, April 2012.

[168] Michael Scott Richards. *Quantitative three dimensional elasticity imaging*. PhD thesis, Boston University, 2007.

[169] Hassan Rivaz, Emad M. Boctor, Michael Choti, Gregory D. Hager, and others. Real-time regularized ultrasound elastography. *Medical Imaging, IEEE Transactions on*, 30(4):928–945, 2011.

[170] Bangti Jin. A variational Bayesian method to inverse problems with impulsive noise. *Journal of Computational Physics*, 231(2):423–435, 2012.

[171] Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. Computer Model Calibration Using High-Dimensional Output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

[172] Maria J. Bayarri, James O. Berger, Rui Paulo, Jerry Sacks, John A. Cafeo, James Cavendish, Chin-Hsu Lin, and Jian Tu. A framework for validation of computer models. *Technometrics*, 49(2):138–154, 2007.

[173] Mark Strong and Jeremy E. Oakley. When is a model good enough? Deriving the expected value of model improvement via specifying internal model discrepancies. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):106–125, 2014.

[174] K. Sargsyan, H. N. Najm, and R. Ghanem. On the statistical calibration of physical models. *International Journal of Chemical Kinetics*, 47(4):246–276, 2015.

[175] Andrew Gelman and others. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534, 2006.

[176] M. Honarvar, R. S. Sahebjavaher, S. E. Salcudean, and R. Rohling. Sparsity regularization in dynamic elastography. *Physics in medicine and biology*, 57(19):5909, 2012.

[177] Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[178] Nicolas Dobigeon and Jean-Yves Tourneret. Bayesian orthogonal component analysis for sparse representation. *Signal Processing, IEEE Transactions on*, 58(5):2675–2685, 2010.

[179] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Ieee Transactions on Information Theory*, 52(2):489–509, February 2006.

[180] D. P. Wipf and B. D. Rao. Sparse Bayesian learning for basis selection. *Ieee Transactions on Signal Processing*, 52(8):2153–2164, August 2004.

[181] Tan Bui-Thanh, Omar Ghattas, and David Higdon. Adaptive Hessian-Based Nonstationary Gaussian Process Response Surface Method for Probability Density Approximation with Application to Bayesian Solution of Large-Scale Inverse Problems. *Siam Journal on Scientific Computing*, 34(6):A2837–A2871, 2012.

[182] RB Muirhead. *Aspects of Multivariate Statistical Theory*. 1982.

[183] Johnathan M. Bardsley, Daniela Calvetti, and Erkki Somersalo. Hierarchical regularization for edge-preserving reconstruction of PET images. *Inverse Problems*, 26(3):035010, March 2010.

[184] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, December 2013.

[185] Arthur Cayley. Sur quelques proprits des dterminants gauches. *Journal fr die reine und angewandte Mathematik*, 32:119–123, 1846.

[186] J. Barzilai and Jm Borwein. 2-Point Step Size Gradient Methods. *Ima Journal of Numerical Analysis*, 8(1):141–148, January 1988.

[187] Ap Dempster, Nm Laird, and Db Rubin. Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1):1–38, 1977.

[188] Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288, 1994.

[189] S. R. Arridge, J. P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen. Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Problems*, 22(1):175, 2006.

[190] Jari Kaipio and Erkki Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007.

[191] A. Samani and D. Plewes. A method to measure the hyperelastic parameters of ex vivo breast tissue samples. *Physics in Medicine and Biology*, 49(18):4395–4405, September 2004.

[192] Joseph J. O'Hagan and Abbas Samani. Measurement of the hyperelastic properties of 44 pathological ex vivo breast tissue samples. *Physics in Medicine and Biology*, 54(8):2557–2569, April 2009.

[193] M. Mooney. Theory of large elastic deformation. *Journal of Applied Physics*, 11:582–592, 1940. INSPEC:1940A02892.

[194] Rs Rivlin. Large Elastic Deformations of Isotropic Materials .4. Further Developments of the General Theory. *Philosophical Transactions of the Royal Society of London Series a-Mathematical and Physical Sciences*, 241(835):379–397, 1948.

[195] Juan C. Simo and Robert L. Taylor. Quasi-Incompressible Finite Elasticity in Principal Stretches - Continuum Basis and Numerical Algorithms. *Computer Methods in Applied Mechanics and Engineering*, 85(3):273–310, February 1991.

[196] Markus Schoeberl. *Comparison of Different Optimization Algorithms for Nonlinear Inverse Problems in Biomechanics*. Master thesis, Technical University of Munich, 2013.

[197] I.M. Franck and P.S. Koutsourelakis. Multimodal, high-dimensional, model-based, Bayesian inverse problems with applications in biomechanics. *Journal of Computational Physics*, 329:91–125, January 2017.

[198] Simon Chatelin, Miguel Bernal, Thomas Deffieux, Clment Papadacci, Patrice Flaud, Amir Nahas, Claude Boccara, Jean-Luc Gennisson, Mickael Tanter, and Mathieu Pernot. Anisotropic polyvinyl alcohol hydrogel phantom for shear wave elastography in fibrous biological soft tissue: a multimodality characterization. *Physics in Medicine and Biology*, 59(22):6923, 2014.

[199] Qianqian Fang, Richard H. Moore, Daniel B. Kopans, and David A. Boas. Compositional-prior-guided image reconstruction algorithm for multi-modality imaging. *Biomedical optics express*, 1(1):223–235, 2010.

[200] Jeremie Fromageau, Jean-Luc Gennisson, Cedric Schmitt, Roch L. Maurice, Rosaire Mongrain, and Guy Cloutier. Estimation of polyvinyl alcohol cryogel mechanical properties with four ultrasound elastography methods and comparison with gold standard testings. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 54(3):498–509, 2007.

[201] Tony Lelivre, Gabriel Stoltz, and Mathias Rousset. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

[202] I. Bilionis and Phaedon-Stelios Koutsourelakis. Free energy computations by minimization of KullbackLeibler divergence: An efficient adaptive biasing potential method for sparse representations. *Journal of Computational Physics*, 231(9):3849–3870, 2012.

[203] Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[204] Steven N. MacEachern and Peter Mller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

[205] Daniela Calvetti and Erkki Somersalo. Hypermodels in the Bayesian imaging framework. *Inverse Problems*, 24(3):034013, 2008.

[206] Michalis Titsias and Miguel Lzaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979, 2014.

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

[208] Evgeny Gladilin and Roland Eils. Nonlinear elastic model for image registration and soft tissue simulation based on piecewise St. Venant-Kirchhoff material approximation. volume 6914, pages 69142O–69142O–9, 2008.

[209] Igor Yanovsky, Carole Le Guyader, Alex Leow, Arthur Toga, Paul Thompson, and Luminita Vese. Unbiased volumetric registration via nonlinear elastic regularization. In *2nd MICCAI Workshop on Mathematical Foundations of Computational Anatomy*, 2008.

[210] Tengxiao Liu, Olalekan A. Babaniyi, Timothy J. Hall, Paul E. Barbone, and Assad A. Oberai. Noninvasive In-Vivo Quantification of Mechanical Heterogeneity of Invasive Breast Carcinomas. *PLOS ONE*, 10(7):e0130258, July 2015.

[211] Radu Dobrescu. Diagnosis of Breast Cancer from Mammograms by Using Fractal Measures. *International Journal of Medical Imaging*, 1(2):32, 2013.

[212] Rangaraj M. Rangayyan, Nema M. El-Faramawy, JE Leo Desautels, and Onsy Abdel Alim. Measures of acutance and shape for classification of breast tumors. *IEEE Transactions on medical imaging*, 16(6):799–810, 1997.

[213] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[214] Yanyan He and Dongbin Xiu. Numerical strategy for model correction using physical constraints. *Journal of Computational Physics*, 313:617–634, 2016.

[215] Pierre Ladeveze and Dominique Leguillon. Error estimate procedure in the finite element method and applications. *SIAM Journal on Numerical Analysis*, 20(3):485–509, 1983.

[216] Marc Bonnet and Wilkins Aquino. Three-dimensional transient elastodynamic inversion using an error in constitutive relation functional. *Inverse Problems*, 31(3):035010, 2015.

[217] Olivier Allix, Pierre Feissel, and Hong Minh Nguyen. Identification strategy in the presence of corrupted measurements. *Engineering Computations*, 22(5/6):487–504, 2005.

[218] Pierre Feissel and Olivier Allix. Modified constitutive relation error identification strategy for transient dynamics with corrupted data: the elastic case. *Computer methods in applied mechanics and engineering*, 196(13):1968–1983, 2007.

[219] Biswanath Banerjee, Timothy F. Walsh, Wilkins Aquino, and Marc Bonnet. Large scale parameter estimation problems in frequency-domain elastodynamics using an error in constitutive equation functional. *Computer methods in applied mechanics and engineering*, 253:60–72, 2013.

[220] Manuel I. Diaz, Wilkins Aquino, and Marc Bonnet. A modified error in constitutive equation approach for frequency-domain viscoelasticity imaging using interior data. *Computer methods in applied mechanics and engineering*, 296:129–149, 2015.

[221] Pierre Ladevze, Jean-Pierre Pelle, Frederick F. Ling, Ernest F. Gloyna, and William Howard Hart. *Mastering calculations in linear and nonlinear mechanics*. Springer, 2005.

[222] Ludovic Chamoin and Pedro Dez. *Verifying Calculations-Forty Years On: An Overview of Classical Verification Techniques for FEM Simulations*. Springer, 2015.

[223] David M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595, 1998.

[224] Hani Eskandari, Septimiu E. Salcudean, Robert Rohling, and Jacques Ohayon. Viscoelastic characterization of soft tissue from dynamic finite element models. *Physics in medicine and biology*, 53(22):6569, 2008.

[225] C. M. A. Vasques and J. Dias Rodrigues. *Vibration and structural acoustics analysis: Current research and related technologies*. Springer Science & Business Media, 2011.

[226] M. H. Kargarnovin, D. Younesian, D. J. Thompson, and C. J. C. Jones. Response of beams on nonlinear viscoelastic foundations to harmonic moving loads. *Computers & Structures*, 83(23):1865–1877, 2005.

[227] Stanislav Pyatykh and Jrgen Hesser. Salt and pepper noise removal in binary images using image block prior probabilities. *Journal of Visual Communication and Image Representation*, 25(5):748–754, 2014.

[228] Athanasios Karamalis, Wolfgang Wein, Tassilo Klein, and Nassir Navab. Ultrasound confidence maps using random walks. *Medical image analysis*, 16(6):1101–1112, 2012.

[229] Sebastian Kehl and Michael W. Gee. Calibration of parameters for cardiovascular models with application to arterial growth. *International Journal for Numerical Methods in Biomedical Engineering*, 2016.

[230] Jeremy G. Siek, Lie-Quan Lee, and Andrew Lumsdaine. *Boost Graph Library: User Guide and Reference Manual, The*. Pearson Education, 2001.

[231] Steve A. Maas, Benjamin J. Ellis, Gerard A. Ateshian, and Jeffrey A. Weiss. FEBio: finite elements for biomechanics. *Journal of Biomechanical Engineering*, 134(1):011005, January 2012.