Robust Bayesian inference via optimal transport misfit measures: applications and algorithms

by

Andrea Scarinci
B.Sc.–M.Eng., Aerospace Engineering, Politecnico di Torino (2013)
S.M., Aeronautics and Astronautics, Massachusetts Institute of Technology (2017)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Aerospace Computational Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2022

Author .....
Department of Aeronautics and Astronautics
April 8, 2022

Certified by .....
Youssef M. Marzouk
Professor of Aeronautics and Astronautics
Thesis Supervisor

Certified by .....
Michael Fehler
Senior Research Scientist, Earth Resources Laboratory
Committee Member

Certified by .....
Ruben Juanes
Professor, Department of Civil and Environmental Engineering
Committee Member

Accepted by .....
Jonathan P. How
R. C. Maclaurin Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Robust Bayesian inference via optimal transport misfit measures: applications and algorithms

by

Andrea Scarinci

Submitted to the Department of Aeronautics and Astronautics
on April 8, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Aerospace Computational Science

Abstract

Model misspecification constitutes a major obstacle to reliable inference in many problems. In the Bayesian setting, model misspecification can lead to inconsistency as well as overconfidence in the posterior distribution associated with any quantity of interest, i.e., under-reporting of uncertainty.

This thesis develops a Bayesian framework to reduce the impact of a type of model misspecification arising in inference problems involving time series data: unmodeled time warping between the observed and modeled data. Inference problems involving dynamical systems, signal processing, and more generally functional data can be affected by this type of misspecification. Inverse problems in seismology are an important example of this class: inaccuracies in characterizing the complex, spatially heterogeneous propagation velocities of seismic waves can lead to error in their modeled time evolution. Data are insufficient to constrain these propagation velocities, and therefore we instead seek robustness to model error. Instrumental to our approach is the use of *transport–Lagrangian* (TL) distances as loss/misfit functions: such distances can be understood as "graph-space" optimal transport distances, and they naturally disregard certain features of the data that are more sensitive to time warping. We show that, compared to standard misfit functions, they produce posterior distributions that are both less biased and less dispersed.

In particular, we use moment tensor inversion, a seismic inverse problem, as our primary motivating application and demonstrate improved inversion performance of the TL loss—by a variety of statistical and physical metrics—for a range of increasingly complex inversion and misspecification scenarios. At the same time, we address several broader methodological issues. First, in the absence of a tractable expression for a TL-based likelihood, we construct a consistent prior-to-posterior update using the notion of a Gibbs posterior. We then compare the impact of different loss functions on the Gibbs posterior through a broader exploration of what constitutes "good" inference in the misspecified setting, via several statistical scoring rules and rank

statistics, as well as application-specific physical criteria. In an effort to link our generalized (Gibbs) Bayesian approach to a more traditional Bayesian setting, we also conduct an analytical and numerical investigation of statistical properties of the transport-Lagrangian distance between random noisy signals.

As a complement to Bayesian inversion, we also demonstrate the utility of optimal transport distances for frequentist regression. We study the linear regression model with TL loss, describe the geometry of the associated mixed-integer optimization problem, and propose dedicated algorithms that exploit its underlying structure. We then compare TL linear regression with classical linear regression in several applications.

Finally, we discuss potential generalizations of TL distances to include the notion of "shape" through time series embeddings, as well as possible extensions of the proposed framework to other forms of model misspecification.

Thesis Supervisor: Youssef M. Marzouk
Title: Professor of Aeronautics and Astronautics

Acknowledgments

While I was in high school, writing my final-year graduation project, humbly titled "*Il caso - The chance*", probability simply meant the mathematics of rolling the dice and, perhaps, a couple of medieval Italian poems. Little did I know, I would have ended up writing a thesis on it. How much chance there was in taking a communication class, and meeting what was soon to become my PhD advisor, is a tough question I will leave for the future members of the UQlab to answer (or quantify!). What is instead certain is that hardly would I have embarked on such a major voyage, if I wasn't sure about the guide I was going to get. A great musician in our group once said: "Before picking an instrument to play, you should pick a good teacher". I sure picked the right one. Thank you Youssef for all these years of mentorship, friendship and fascinating conversations. I hope we will be in touch for many years to come.

Irreplaceable part of this journey, I cannot forget Michael Fehler, who, besides being my favorite seismologist, could not have been a better advisor in my first approach to the world of earthquakes and misspecified velocity models. In this regard, I must also mention Ruben Juanes for being a thoughtful and supporting committee member. Thank you as well to Ludovic Metivier and Alice Cicirello for patiently reading my thesis and, Alice, for an unforgettable two week project at Oxford university (and beyond).

On the other side of the world, a big thank you also goes to my collaborators in Dahran: Umair Bin Waheed, SanLinn Kaka and Ben Dia. You enriched this thesis with your different contributions and perspectives, and made our trips to Saudi Arabia a lasting memory.

Finally, a PhD is never complete without the daily companionship and friendship of labmates, officemates and staff. So here I try to thank you all, name by name, hoping not to forget anyone: Ricardo, Ben, Feng-Yi, Sven (spelled the right way), Michael, Rebecca, Alessio, Olivier, Daniele, Shaohong, Pablo, Rémi, Jouni (the great musician), Elizabeth, Chaitanya, Chi, Zheng, Aimee, Paul Baptiste, Matt, Josh, Jean,

Beth, Lucio, Siheng and Yu. You all made me feel home and, for this reason, a bit sad to having to say goodbye now. *A presto!*.

—— The author would like to acknowledge support from the King Fahd University of Petroleum and Minerals (KFUPM), College of Petroleum Geosciences (CPG), Global Partnership Program with MIT. ——

Contents

List of Figures

List of Tables

Chapter 1

Motivation and outline

1.1 Introducing the problem: model misspecification, misfit measures, and time series

Model error or misspecification is a determining factor in the quality of the solution of an inverse problem. Consistency of the Bayesian framework, in particular, heavily relies on the characterization of the observation error and modeling of the underlying physical phenomenon. This thesis focuses on a particular kind of model misspecification that arises when dealing with time series data or data that requires some kind of discretization over time or space (e.g., images). In such cases a vector or a matrix contains the intensity or amplitude of the object of interest at specific time or space coordinates and it is assumed that this mapping is consistent between the observed and modeled data. In other words, it is assumed that no kind of warping is necessary to map the discretized point of the modeled data to the observed one (or vice versa). Under this premise, the use of ℓ_p norms as misfit functions is a natural choice: the value observed at time-index i is mapped (compared) to that of the modeled signal at the same time index i. An ℓ_2-norm misfit function is particularly common, as it matches the notion of additive Gaussian noise. In reality, it may often occur that portions of the time series are anticipated or delayed with respect to the model predictions

due to various kinds of misspecification: from incorrect modeling of observational error to deficient modeling of the underlying physics. When this occurs the use of ℓ_p norms can have unintended consequences: two signals or images that may look very similar in "shape" to the human observer may look far apart under an ℓ_p norm due to misalignment in the time-space coordinates. These norms, in fact, practically ignore the relationship between the different coordinate values and treat the time series as a collection of uni-dimensional data points.

In the context of seismic waveform inversion, for example, this kind of misspecification is particularly relevant when it comes to modeling the propagation velocity $\mathbf{V}$ of a seismic wave. Due to the extreme difficulties in characterizing the subsurface medium (e.g., different rock types, three-dimensional spatial heterogeneities) any velocity model is generally approximate and inaccurate. Mischaracterization of $\mathbf{V}$, however, can impact one's ability to infer other quantities of interest such as the hypocenter $\mathbf{x}$ and the moment tensor $\mathbf{m}$ (focal mechanism) of a seismic event. In deterministic full-waveform-inversion this often results in the well known phenomenon of cycle-skipping, which traps optimizers in local minima [47]. In the Bayesian setting, model misspecification can lead, in a worst case scenario, to overconfidence in the posterior distribution, i.e., under-reporting of uncertainty [56, 72].

The most direct approach to mitigating the impact of model misspecification is to introduce better physical models (when feasible) or improved statistical discrepancy models. These approaches, however, typically increase computational cost and may compromise parameter identifiability. As an example, in moment tensor inversion, using a simple layered-medium model for the propagation velocity can be orders of magnitude less expensive to run than a fully three-dimensional elastic wave propagation model. Moreover, such sophisticated models are typically not available for the majority of sites, and data for learning the velocity jointly with the focal mechanism in such a three-dimensional setting may be entirely unavailable, and confounded with the estimation of the focal mechanism itself.

In this thesis we instead investigate the benefits of using an alternative, optimal

transport (OT) based, misfit function to measure discrepancies between observed and model predicted data. Recent literature has demonstrated the applicability of OT-based distances to seismic imaging problems in a deterministic setting [42, **?**, 87, 88, 154]. In this context, OT has been shown to produce drastic reductions in the non-convexity of the objective function, especially when compared to ℓ_p distances. A more convex misfit function also implies a more robust solution to the inverse problem when subject to uncertainties in the input parameters. Rigorous mathematical treatment [43] has in fact shown that 1-D quadratic Wasserstein distances (a subset of OT distances) are convex functions with respect to dilation and translation when applied to probability density functions. In order for this to remain valid for generic signals as well, it is however necessary to normalize and positivize them accordingly.

This last requirement introduces data transformations that are not typically justifiable within the physics of the problem. We therefore propose to focus on a particular case of Wasserstein distance that does *not* require signal positivation and normalization, and therefore makes it more suitable to deal with seismic waves. This distance is referred to as the transport–Lagrangian (TL) distance [128, 129, 73] and can be interpreted as the result of solving an optimal transport (OT) problem between the graphs of two functions.

While the benefits of using this kind of distance have been already explored in a number of deterministic inverse problems and applications [128], including seismology [87, 88], in this thesis we formulate and explore its integration within a fully Bayesian framework. Within this setup, we interpret the TL distance as a tool to tackle a broader issue than the cycle-skipping issue highlighted in the deterministic literature on full waveform inversion (FWI). More precisely, we look at the TL distance as a data "feature extractor" that deliberately disregards information not relevant to the inference of a particular quantity of interest, minimizing the impact of uncertainties in the model.

1.2 Thesis outline and contributions

The first step in the development of a coherent Bayesian procedure is to establish a statistical model for the phenomenon of interest and then to derive the associated likelihood function. We maintain the classical additive Gaussian noise setup, which normally corresponds to an ℓ_2 norm misfit in the exponent of the Gaussian probability density function (PDF). Instead, however, we introduce the TL distance as our misfit statistic. For such a statistic, it is not straightforward to characterize the conditional distribution of the distance given a particular value of model parameters and a model for the additive Gaussian noise. We discuss some results on this topic from recent literature [4, 16, 94, 37] and propose the use of so-called Gibbs posteriors, and their interpretation in the misspecified context. We also derive with more detail certain statistical properties and behaviors of properly defined TL-based likelihood function, including a closed form expression for it.

Once the framework is defined, it is important to choose some criteria to quantitatively assess its performance. To this purpose, we propose a number of quantitative metrics to characterize the kind and degree of improvement introduced by the TL distance. We compare the resulting posterior distributions to those obtained by using classic ℓ_2-based Bayesian frameworks. We emphasize that there is no unique way to establish in what respect one posterior distribution is better than another, and if so, how this depends on the specific use that that the analyst intends to make of it. For this reason, we look at two different scoring rules that exist in the literature. Continuous rank probability scores (CRPS) [52, 51] effectively capture two important qualities a posterior distribution needs to have in order to be used as a practical forecaster: be sufficiently localized (i.e., low variance) and contain the true value of the quantity of interest within its support, preferably in high-probability regions (i.e., low bias). The perfect forecaster would therefore be a delta distribution located at the true value of the quantity of interest θ: $\delta(\theta)_{\theta=\theta_{\text{true}}}$. Aside from CRPS, we also discuss ways of checking the self consistency of the inference procedure, specifically in the case of the TL-based likelihood function. For this purpose we focus on recent

literature proposed around the concept of rank histograms [126, 27]. These are checks on the frequentist behavior of Bayesian credible intervals and entirely *within-model* assessment tools. The objective is to verify whether the inference procedure shows any bias in reporting uncertainty around certain regions of the parameter space.

As the main testbed for the proposed framework, we construct several instantiations of the seismic moment tensor estimation problem in which the data generating process relies on a different velocity model $\mathbf{V}$ than the one used for modeling predicted waveforms. Our empirical studies exhibit increasing levels of complexity and realism. Overall, we will show that the TL-based likelihood is less sensitive to the nuisance effect introduced by the misspecified $\mathbf{V}$ and allows for the construction of more informative posterior distributions on $\mathbf{m}$.

Outside the domain of Bayesian inversion, misfit measures for time series and images also play a central role in classic linear regression problems. These as well can be affected by the misspecification issues described above and can benefit from the use of alternative misfit functions. In this thesis we formulate a linear regression method that instead uses the transport-Lagrangian (TL) distance as the objective function to be minimized. The associated optimization problem exhibits an increased complexity over the traditional least-squares setting, since it allows to optimize not only over the regression coefficients, but also over the amount of transport to perform between modeled and observed data. In other words it combines a continuous quadratic program with a discrete optimal assignment problem. We will propose a dedicated algorithm and test it on a number of applications.

All of the content described above is articulated and detailed within the following chapters:

- Chapter 2 contains a more detailed and precise definition of the problem of **model misspecification**, coupled with background **literature** on common approaches to mitigate it. The principal **motivating application** for the thesis, i.e., moment tensor inversion, is also presented, with an emphasis on velocity model misspecification;

- Chapter 3 presents the first main contribution of this thesis, i.e., the definition of a consistent **Bayesian** inference **framework** to incorporate **optimal transport distances** as robust misfit measures. The transport-Lagrangian distance in particular is defined and associated algorithms for its computation are discussed. The second part of this chapter is dedicated to answering the question of how to quantitatively evaluate or **score posterior distributions**. A number of criteria are discussed as well as their advantages and disadvantages;

- Chapter 4 is dedicated to testing the proposed framework on a realistic velocity model for moment tensor inversion: the **SEG-EAGE Overthrust model**. Possible implications of the obtained results for some problems of geophysical nature are also discussed;

- Chapter 5 is dedicated to characterizing a **TL-based likelihood function** associated to an additive Gaussian noise model. Some asymptotic results are proposed together with a closed form expression. Particular care is also taken in describing the geometry of the statistical model;

- Chapter 6 presents some algorithmic considerations and formulations for a deterministic, **TL-based linear regression** problem. Applications to demonstrate the viability and usefulness of this approach are also described.

In the last chapter some conclusive remarks are gathered together with an outlook on possible extensions and generalization of the proposed framework to other types of model misspecification.

Chapter 2

Model misspecification: background

2.1 Model misspecification in Bayesian inference and inverse problems

Model misspecification in general can have a permanent impact on the ability to perform accurate inference. In Bayesian inference this can manifest itself in the prior not including the truth, or not placing sufficient probability on it. More often the likelihood (which in this view includes the forward model) may not reflect the true data-generating process. For *finite-dimensional* parameters, prior distributions are perhaps less sensitive to this issue, since an infinite amount of data could in principle correct any belief about the parameter values, unless the support of the prior does not include the true parameter values.[1] But consistency of the statistical model for the data (as encapsulated in the likelihood function) with the true data-generating process is essential to achieving meaningful results.

A vast and growing body of literature exists on model misspecification and strategies for how to perform robust inference. In this chapter we will briefly recall what is meant by model misspecification and discuss some of the most common approaches to

[1]This aspect is more complex and subtle in the infinite-dimensional setting of Bayesian non-parametrics; see, e.g. [32, 99]. We will avoid these complexities and work only in the setting of finite-dimensional parameters here.

make Bayesian inference more robust.

2.1.1 The Bernstein–von Mises theorem

For a consistent Bayesian update, it is generally assumed that the distribution of the data belongs to the family of parameterized distributions defined by the model. More formally, if $\{y_i\}_{i=1}^n$ is a sequence of i.i.d. random variables each with density $g(y_i)$ (the true data distribution, generally unknown) and $\{f(y_i|\theta),\ \theta \in \Theta\}$ is a family of parameterized densities to approximate $g(y_i)$, we say that the model is well-specified if there exists a $\theta_0 \in \Theta$ such that $g(y_i) = f(y_i|\theta_0)$. Under such premises (and some additional technical conditions), the standard Bernstein–von Mises theorem holds [49, 135]. This result ensures that the posterior distribution, asymptotically in the size of the data set n, becomes Gaussian and centered around the true parameter value θ_0. The scale of the posterior covariance shrinks, asymptotically at a $1/n$ rate, and posterior credible intervals are guaranteed to have good frequentist coverage.

In contrast, when the model is misspecified, i.e., $g(y_i) \neq f(y_i|\theta)$ for any choice of $\theta \in \Theta$, the posterior distribution will, asymptotically in n, become Gaussian but centered around a value θ^* which is [72]:

$$\theta^* = \arg\min_{\theta\in\Theta} \mathcal{D}_{\mathrm{KL}}(g(\cdot)\,||\,f(\cdot|\theta)). \tag{2.1}$$

where:

$$\mathcal{D}_{\mathrm{KL}}(g(\cdot)\,||\,f(\cdot|\theta)) = \int_{-\infty}^{\infty} g(\cdot)\log\left(\frac{g(\cdot)}{f(\cdot|\theta)}\right)d\cdot$$

is the Kullback-Leibler divergence between two probability distributions. Minimizing the KL divergence will not ensure that the model $f(y|\theta^*)$ will be able to reproduce the data. Moreover, the KL distance does not necessarily have a unique minimizer over Θ. The posterior covariance will still shrink towards zero as new data is incorporated, and the posterior distribution may therefore provide a misleading—in particular, overly confident—characterization of the uncertainty in the problem.

2.1.2 Common approaches to mitigating model misspecification

A certain amount of research has been conducted on how to make inference more robust to model misspecification [149, 65, 24]. The classical approaches can be categorized in two threads [119]:

- Better/more articulated physical modeling: either in terms of the actual physics of the phenomenon or in terms of model selection/extension [107, 108, 21, 113];
- Better or more robust statistical modeling of the data: for instance, moving beyond Gaussian additive noise, as in $y(t_i) = u(\theta, t_i) + \epsilon_i$ for some parameters θ, deterministic forward model u, Gaussian random variable ϵ_i, and covariate values t_i, $i = 1 \dots n$. Along these lines, the influential modeling approach of [71] argues for the addition of a Gaussian process discrepancy model $\delta(t)$ to the relationship above:

$$y(t_i) = u(\theta, t_i) + \delta(\phi, t_i) + \epsilon_i \tag{2.2}$$

  where ϵ_i remains a Gaussian noise that represents measurement error and the term $\delta(\phi, t)$ is a Gaussian process, indexed by t, aimed at statistically modeling additional mismatch or *discrepancy* between the observed data and model predictions [71]. Here ϕ are additional parameters describing the Gaussian discrepancy process, not necessarily related to θ.

Both strategies present advantages and disadvantages. Using more complex physical models can of course increase the chances of matching the observations; however this often comes at increased computational cost and/or parameter identifiability issues. The same can be said for the term $\delta(\phi, t)$, where the additional parameters ϕ need to be estimated. An additional concern regarding this approach is that, by calibrating ϕ through the data, it may be difficult to discern whether the term δ is only compensating for missing statistical modeling or it is instead acting as a compensator

for the physical model itself. Of similar flavor, although not intended to better capture the statistical nature of the phenomenon, are methods that go under the name of *model space extension or source extension*. In this case, additional degrees of freedom are introduced at the beginning of the inversion to help fit the data. These additional (and artificial) degrees of freedom are iteratively converged to 0 along the inversion process. This approach has been of particular interest in seismic inverse problems [125] (see next section for more details on this specific application). The main risk of such approaches is that while the observed data may be better fitted, the prediction capabilities outside the dataset itself may be extremely poor.

Some more recent approaches revolve around the concept of *coarsening* [93]. The key idea is to modify the standard Bayesian approach to introduce a posterior distribution obtained by conditioning not on the event that the data are generated by the model distribution, but instead on some measure of discrepancy between the observed data $y_{1:n}$ and model-predicted data $y^{\theta}_{1:n}$. In formulae:

$$p(\theta \,|\, \mathcal{D}(y_{1:n}||y^{\theta}_{1:n}) < \epsilon) \propto p(\theta)\mathbb{P}\big[\mathcal{D}(y_{1:n}||y^{\theta}_{1:n}) < \epsilon\big], \tag{2.3}$$

where $y_{1:n}$ are i.i.d. data, $p(\theta)$ is a prior probability density, and $\mathcal{D}$ is a generic measure of discrepancy between the two (empirical) distributions of the data. When the discrepancy measure is chosen to be an empirical approximation of the Kullback–Leibler (KL) divergence, the posterior distribution defined in (2.3) can be approximated by:

$$p(\theta \,|\, \mathcal{D}^{n}_{\mathrm{KL}}(y_{1:n}||y^{\theta}_{1:n}) < \epsilon) \propto p(\theta) \prod_{i=1}^{n} f(y_i|\theta)^{\xi(n,\epsilon)}. \tag{2.4}$$

The parameter $\xi(n, \epsilon) \in (0, 1)$ effectively acts as a coarsener and should be chosen to depend both on the number of samples and ϵ. Intuitively, the coarsening construction produces the following effect: as long as some discrepancy is present between the observed and model-predicted data, the posterior distribution will not concentrate around a specific value, even with an infinite amount of data. This approach avoids the undesirable posterior concentration described by the Bernstein–von Mises theorem

under misspecification [72]. Although a technique is presented in [93] to choose $\xi(n, \epsilon)$ in systematic way, the procedure still involves some discretionary aspects. A similar "coarsening" of the posterior distribution is described in [96], where the shrinking covariance matrix (with asymptotic scale $1/n$) deriving from the asymptotically normal behavior of the posterior distribution [72] is replaced by a covariance matrix that takes into account the discrepancy between the predicted and observed data. This covariance, called the "sandwich," does not shrink even with infinite data as long as a discrepancy between predictions and observations is present.

Another line of research to address model misspecification has its roots in decision theory [96, 148, 55]. Here robustness to model misspecification is assessed by the means of a minimax rule. In [148] a loss function $\mathcal{L}(\theta)$ is defined with the model parameter θ as an argument. A posterior distribution $p(\theta|y)$ is calculated given the best available information in terms of modeling, data, and prior distributions. Subsequently, a set Γ_C of distributions $p_C(\theta)$ is defined such that they all lie (in a KL sense) within a radius C of the calculated posterior $p(\theta|y)$:

$$\Gamma_C = \{p(\theta) : \mathcal{D}_{\mathrm{KL}}(p(\theta) \,\|\, p(\theta|y)) \leq C\}. \tag{2.5}$$

An upper bound is then calculated for the expected loss over all possible distributions contained in Γ_C:

$$p^{\sup}_{\Gamma_C} = \sup_{p(\theta)\in\Gamma_C} \mathbb{E}_p[\mathcal{L}(\theta)]. \tag{2.6}$$

At this point, depending on whether the value of the maximum expected loss is acceptable or not, the analyst can decide whether to improve the model further. The definition of the loss function itself, and the maximum acceptable radius in which the perturbed posterior can lie, are both the results of choices that the analyst must make *a priori*, based on the specific scope of the study.

In the same fashion one can look for the minimizer of the expected loss, and thus end up with a pair of distributions $\{p^{\sup}_{\Gamma_C}, p^{\inf}_{\Gamma_C}\}$ that characterize the robustness of the posterior $p(\theta|y)$ for a given loss. This methodology is interesting as it presents a general

framework under which many existing techniques (including the likelihood-coarsening technique discussed above) can be incorporated. The definition of the loss function as well as the set Γ_C are flexible enough to incorporate the features the analyst cares about.

Also based on decision-making strategies is the recent concept of "safe Bayes" [55]. Here, the analyst defines a set $\mathcal{P}^*$ of credible distributions on θ, according to some criteria of interest. Informally, $\mathcal{P}^*$ is a set that is subjectively believed to contain the *true* posterior (i.e., the posterior on θ that would be obtained with a well-specified model) $p_{\text{true}}(\theta)$. After collecting some data, a standard (and in general misspecified) Bayesian posterior distribution $p(\theta|y_{1:n})$ is obtained. This distribution, called the *pragmatic distribution*, can be deemed as "safe" for predicting θ if the following condition holds:

$$\forall p \in \mathcal{P}^*, \qquad \mathbb{E}_{\theta\sim p}[\theta] = \mathbb{E}_{\theta\sim p(\theta|y_{1:n})}[\theta]. \tag{2.7}$$

In words, this means the posterior update does not alter or bias the subjective knowledge about θ in a any systematic way. While the notion of robustness is here clearly established, the notion of "credible set" of posteriors relies on subjective judgement by the analyst.

2.1.3 An alternative perspective

In this thesis, by taking the seismic inverse problem as a reference application, we take a different perspective than those offered by the methodologies discussed so far. Given that the model complexity of seismic wave propagation is already high enough, more sophisticated modeling would not be the path to follow. Coarsening would make inference more robust to model misspecification while keeping the complexity of the model fixed. However, it would do so in a generalized fashion, by increasing the variance of the posterior distribution without taking into account whether, for at least a subset of the quantities of interest, it is still possible to capture the true parameter

value. Decision theoretic frameworks, although theoretically sound, require a number of assumptions to define with respect to what the posterior shall be considered robust to. How to build credible sets and loss functions can in fact be a challenging problem by itself, and the result consists again in adding more uncertainty to the posterior distribution, rather than directly tackling model misspecification.

We will describe our approach to model misspecification beginning in Chapter 3 (noting it can be combined with the ones just described). Before that, we turn to some review of the seismic inverse problem, which will make the preceding points clearer.

2.2 An example: full waveform inversion and incorrect velocity models

2.2.1 The seismic inverse problem

A major goal in seismology is to understand how seismic waves propagate through a given subsurface medium (*forward problem*). Parallel to this is the so-called *seismic inverse problem*, which relates the observed seismic displacements (typically recorded by seismograms on the Earth's surface) to their source (earthquake). Characterizing earthquakes provides a better understanding of the earth processes and is of particular interest in the oil and gas as well as geothermal industries, where small earthquakes are artificially induced by activities such as mining, fluid injection and oil production. At least two main subproblems can be identified within seismic inversion. The first one aims at reconstructing the structure of the subsurface assuming the hypocenter of the earthquake is correctly localized as well as its time-history appropriately described. In this scenario, typical quantities of interest are velocity models, densities or other elastodynamic properties of the subsurface. Another type of inversion targets instead the characterization of the source. This includes the location of the source, its time-signature and focal mechanism (moment tensor). In this dissertation, we will focus on the second of the two problems just described.

The equation at the heart of this problem is the momentum equation for a three-dimensional elastic continuum:

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \sum_{j=1}^{3} \frac{\partial \tau_{ij}}{\partial x_j} + f_i, \quad i = 1, \ldots, 3, \tag{2.8}$$

where $\rho(\boldsymbol{x})$ is the medium density, $u_i(\boldsymbol{x}, t)$ is the displacement in direction i, $\tau_{ij}(\boldsymbol{x}, t)$ is the ij-th element of the stress tensor, and $f_i(\boldsymbol{x}, t)$ is the body force along direction i. In order to solve the above equation for the displacements u_i it is necessary to relate the stress tensor elements to the u_i-s, via Hooke's law. In particular, for an homogeneous and isotropic medium:

$$\tau_{ij} = \lambda \delta_{ij} \frac{\partial u_k}{\partial x_k} + \mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \tag{2.9}$$

where λ and μ are Lame's parameters and $\delta_{ij} = 1$ for $i = j$ and zero otherwise. Solving (2.8) and (2.9) for the displacement fields $u_i(\boldsymbol{x}, t)$, $i = 1, \ldots, 3$, is generally difficult [116]. One approach to solving this PDE is to express the solution $u_i(\mathbf{x}, t)$ (the displacement in direction i at location $\mathbf{x}$ and time t) in terms of the Green's function $\mathbf{G}_i(\mathbf{x}, \mathbf{x}_s, V, t)$, which is the solution of the PDE at $(\mathbf{x}, t)$ when a unit impulse is applied at $\mathbf{x}_s$ (earthquake or source location) and $t = 0$, and when the velocity model is $V(\mathbf{x})$. The displacement due to a single seismic event at $(\mathbf{x}_s, 0)$ can then be expressed as:

$$u_i(\mathbf{x}, t) = \mathbf{G}_i(\mathbf{x}, \mathbf{x}_s, V, t) \cdot \mathbf{m}^T, \tag{2.10}$$

where $\mathbf{G}_i$ is the Green's function and $\mathbf{m}$ is the *moment tensor*, which represents the force couples that represent an earthquake. In its most general form, $\mathbf{m}$ is a 3×3 symmetric matrix, meaning only six of its elements are independent. This allows one to recast it as a 1×6 vector $\mathbf{m}$ as indicated in (2.10), where $\mathbf{G}_i$ is also a 1×6 vector for any set of input values. Further simplifications and decompositions are possible when the earthquake mechanisms are restricted to be of a particular type (e.g., double couple).

The Green's function $\mathbf{G}_i$ contains, implicitly, all the information relevant to the

seismic phenomenon beyond the source term. This includes quantities such as the location of the earthquake $\mathbf{x}_s$, the density $\rho(\mathbf{x})$ or the propagation velocity fields of the primary and secondary waves, $V(\mathbf{x}) = (V_p(\mathbf{x}), V_s(\mathbf{x}))$ respectively, which in turn can be determined by the characterization of the stiffness tensor when the medium cannot be considered isotropic [3]. The objective of full seismic waveform inversion is that of inferring one or a subset of these quantities of interest, given some observed displacements $y_i(\mathbf{x}, t)$ (normally recorded through seismograms positioned at given locations on the field of interest). A typical choice is to invert for the velocity model. Even though the velocity is in general not homogeneous with respect to $\mathbf{x}$, in most applications it is restricted to assume some fixed values within a certain portion or layer of the terrain of interest, reducing the complexity of the model. For this reason the notation is simplified to $V(\mathbf{x}) = \mathbf{V}$.

In this thesis, our seismic applications will focus on estimating the moment tensor components, while considering all other parameters (particularly $\mathbf{V}$ and $\mathbf{x}_s$) fixed to given values. We will generally refer to the Green's function by making explicit its dependence on the location of the source earthquake and the velocity model only. For any time t, let $\mathbf{u}(t)$ be the vector containing the displacements for each direction and each station/location of interest. If k is the number of stations and we consider all three components of displacement, then $\mathbf{u}(t) \in \mathbb{R}^{3k}$ for any t. The Green's function then becomes a matrix-valued function: for any fixed set of arguments, it is a $3k \times 6$ matrix. We can now write:

$$\mathbf{u}(t) = \mathbf{G}(\mathbf{x}_s, \mathbf{V}, t) \cdot \mathbf{m}^T, \tag{2.11}$$

The Green's function $\mathbf{G}$ is a nonlinear function of $\mathbf{x}_s$ and $\mathbf{V}$. This implies that the objective function of a typical least squares minimization problem for these parameters will most likely be non-convex and that, in a Bayesian setting, the full posterior distribution $p(\mathbf{x}_s, \mathbf{V}, \mathbf{m}|\mathbf{y})$ will be non-Gaussian.

2.2.2 Velocity model uncertainty

Any velocity model is an imperfect representation of the subsurface and cannot properly account for the 3D structure of a region. Inhomogeneities and the difficulty of directly observing the Earth's structure have induced seismologists to find alternative strategies to velocity modeling. How to construct reliable models has been a longstanding issue in seismology to which a definitive answer is yet to be provided [153, 120]. A common approach that we will consider throughout this thesis is to generate model waveforms using a layered medium model (e.g., [34]). This model is often derived from well logs or from some other model of the subsurface, such as one derived from arrival-time tomography [59] or kinematic source representation [111]. Of course, this adds considerable uncertainty to the results of any associated inverse problem and, in general, looking at the effects of layered medium approximations to 3D velocity models is also at present an undeveloped area of research.

Because the propagation velocity of seismic waves impacts the timing at which the waves reach the surface, velocity modeling errors can translate into the type of misspecification outlined in the previous section. As an example, we report in 2-1 a pair of waveforms—i.e., *displacements* $u_i(\mathbf{x}, t)$, for some direction of displacement i and a fixed surface location $\mathbf{x}$ - coming from two different velocity models: the one in blue come from a 3D model and the one in orange come from a 2D layered-medium velocity model built from well logs. It is evident that some kind of warping occurs between the two traces, which are otherwise similar in "shape."

2.2.3 Methods for solving the seismic inverse problem

We conclude this section with a survey of some common waveform inversion methods that have been adopted, mainly in deterministic inversion. While none of these methods specifically aims at tackling the problem of model error, the variety and nature of the proposed approaches convey the complexity of the problem and are symptoms of the issue this thesis tries to solve: model misspecification.

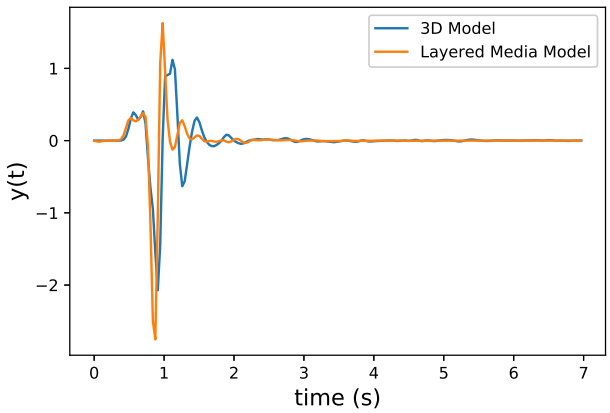


Figure 2-1: Sample waveforms coming from two different velocity models

Local optimization based approaches

Waveform inversion started to become a problem of interest in seismology since the introduction of seismograph networks capable of recording accurate seismic data (both active and passive). The most traditional approach to seismic inversion is non-probabilistic, i.e., the model parameters are recovered by optimizing a misfit function defined over observed $\mathbf{y}(t)$ and model-predicted $\mathbf{u}(t)$ waveforms [141]:

$$\mathcal{M}(\theta) = \mathcal{M}(\mathbf{y}(t), \mathbf{u}(\theta, t)), \tag{2.12}$$

where θ indicates any subset of the parameters of interest to be recovered (the velocity field, earthquake location, moment tensor, density of the media, etc.). However, given the complexity of the forward model and ill-posedness of the inverse problem, the minimization of $\mathcal{M}$ is often performed locally, meaning the solution is sought only in the vicinity ($\Delta\theta = \theta - \theta_0$) of an initial model configuration $\mathbf{u}(\theta_0)$ [141]:

$$\mathcal{M}(\theta) = \mathcal{M}(\theta_0 + \Delta\theta). \tag{2.13}$$

Typically, misfit functions are chosen to be norms, with the squared ℓ^2 norm, as in a least squares problem, being the most popular choice [98]. A number of techniques (Newton, truncated Newton [91], Gauss-Newton, gradient or steepest descent) are

used to solve, under various assumptions, the optimization problem. The ill-posedness of the seismic inverse problem and associated non-convexity of $\mathcal{M}(\theta)$ has motivated further attempts to constrain the number of possible solutions beyond the perturbation/linearization methods just discussed. Regularization of the objective function such as Tikhonov regularizers with various weighting (inverse covariance) matrices are part of the classic repertoire [8, 9, 132]. Recent advances in compressed sensing and the associated idea of randomized data sampling have also motivated approaches to full-waveform inversion that look for for sparse solutions [152, 83, 77, 158].

Alternative measures of misfit have also been proposed such as the ℓ^1 norm, secant and mixed ℓ^1–ℓ^2 norms (Huber penalty) [29], as well as optimal transport based misfit functions [89, 25, 90, 92]. In [122] the authors proposed a general robust algorithmic framework to account for different types of misfit functions and regularization terms. The algorithm relies on quasi-Newton Hessian approximation methods to minimize the misfit function and proximal gradient methods to minimize the regularizing terms.

Of similar flavor are cross-correlation approaches [136] together with deconvolution approaches [80, 145, 58]. In both cases the aim is to minimize the impact of phase traveltime differences, or relative phase shifts, while generally mitigating the well-known phenomenon of cycle skipping [144]. With the same objective it is also worth mentioning methods based on instantaneous phase differences and envelope ratios between observed and synthetic seismograms [18, 109, 79].

Gradient-free methods have also been adopted. In the context of downhole micro-seismic moment tensor inversion [46], a grid-search is performed over plausible event locations and velocity models. The best fitting solution, in a least squares sense, is then identified. In a similar fashion, a large body of literature exists on how to constrain the number of feasible solutions by estimating source-receiver travel times and rejecting solutions that are not compatible (ray-tracing techniques) [116].

Solutions of the problem both in the time and frequency domains have also been attempted [156, 110, 104, 19, 68]. The advantage of using one method over the other depends on several factors such as the type of data available (high/low frequency,

noise level, etc.) as well as what kind of information is recovered in the inversion. In [134], moment tensor inversion is performed through a two-step procedure involving an inversion both in the time and frequency domains, exploiting the linearity of $\mathbf{u}(t)$ in $\mathbf{m}$ when the velocity model and source location are fixed.

Multiscale or multifidelity approaches have also been adopted in the spirit of mitigating the computational complexity and ill-posedness of the problem [104, 22]. Given recent advances in computational power, 3D full waveform inversion have become customary in most applications, especially in industry [104, 90, 147, 139, 138, 57, 78, 155]. From a methodological perspective, however, most of the theory developed I still anchored to 2D models [48].

The main bottleneck to the success of these approaches is often the accuracy of the starting model, and the validity of assumptions underlying various simplifications (e.g., linearization). It is hard to build a valid initial model for optimization, especially when it comes to estimating the velocity $\mathbf{V}$. Starting with a highly misspecified model inevitably leads to bad parameter estimates. This problem is accentuated by the fact that deterministic inversion typically yields only single-point estimates; the uncertainty that surrounds the solution is largely ignored. We argue that a Bayesian framework offers a more complete representation of one's current state of knowledge, and is particularly relevant in a misspecified setting.

Bayesian formulations

An extensive amount of work exists around full waveform inversion performed in a Bayesian framework [81, 36, 53, 115, 107, 60, 157, 118, 67, 48]. As already stated in the introduction, the most common statistical assumption regarding the relation between observations and model predictions is that of additive Gaussian noise. As for the prior distribution, its choice largely depend on the information available to analyst prior to inversion. A central aspect of Bayesian computation in seismic inversion is the computational cost associated with evaluating the posterior distribution and its marginals. Characterizing the moments of the posterior distribution or calculating its

normalization constant is often impossible analytically and involves high dimensional integrals when performed numerically. For iterative methods, the computation of the forward model at each iteration also poses computational challenges. In [115, 31] a number of sampling approaches have been considered, from Metropolis-Hastings Markov chain Monte Carlo (MCMC) to Gibbs sampling and other techniques. An interesting approach to mitigate the dimensionality issue implicit in seismic Bayesian inverse problems is the trans-dimensional Markov chain Monte Carlo sampler implemented in [107]. This method represents a way to compromise between computational cost and modeling accuracy. The cited work relies, however, on the parameterization of the velocity field through wavelets, which represents itself a simplification of the phenomenon. A similar approach for velocity and density recovery is also implemented in [108] with similar results.

Given these challenges, Bayesian approaches to full waveform inversion have usually required some degree of simplification of the problem to make it tractable. In [10] the authors propose a Bayesian approach in which the forward model is linearized in the velocity $\mathbf{V}$. The error and the prior distribution are chosen to be Gaussian, which allows the posterior distribution to be derived analytically. The results of this study show that, given a synthetic data-set, although good agreement is found with the correct solution, a signal to noise ratio of 15 (relatively weak noise) introduces high degrees of uncertainty in the solution. The simplifications introduced are probably partly responsible for these unsatisfactory results. In [127] the authors also linearize the problem, but around the *maximum a posteriori* (MAP) parameter estimate. Despite the simplification, the computational challenge of determining the MAP persists. It involves a nonlinear and non-convex optimization problem and there is no guarantee that the approximation will be good enough. Other Bayesian inversion attempts with linearized models can be found in [60, 54]. An alternative consists in pre-computing the forward model over a grid of possible parameter values [121]. While computationally efficient, this approach poses the question of how fine the grid over the parameter space has to be, and can quickly become impractical when performing multi-parameter inversion.

In [56] the authors attempted a full Bayesian inversion without any simplification of the forward model or linearization around specific solutions. To overcome the computational challenges and increase robustness of the solution, a number of sampling strategies were implemented to exploit conditional linearities and associated Gaussianities in the problem. These include marginal-then-conditional sampling, pre-computing a library of velocity models and source locations, as well as *coarsening* as described in [93]. The results achieved through this implementation are satisfactory when the velocity is known and set to a specific value. As soon as uncertainty is introduced in $\mathbf{V}$, the solutions of the problem exhibit a high degree of instability, indicating model misspecification issues.

More recent approaches have included a proof of concept study for Hamiltonian Monte-Carlo [48] as well as an ensemble Kalman filtering approach applied to full waveform inversion [130].

2.3 Conclusions

In this chapter we described the problem of model misspecification in Bayesian inference both from a theoretical point of view as well in terms of its practical implications in Bayesian inverse problems. We discussed the main strategies proposed so far in the Bayesian literature to avoid over-concentration of posterior distributions. In relation to these methods, we outlined a pathway to a different approach that relies on the choice of specific misfit measures. Finally, we presented the main application of this thesis: moment tensor inversion under misspecified velocity models. We discussed what the misspecification implies in this context and how the problem has been tackled both in the field of deterministic and Bayesian full-waveform-inversion.

Chapter 3

Optimal transport misfit measures for robust Bayesian inference

3.1 Optimal transport distances and time series

3.1.1 Motivation and background

Central to any inverse problem, both in the deterministic and Bayesian framework, is the choice of a misfit function to compare model and predicted data. We have described in the introduction how choosing this metric can play a determining role in performing good inference, especially when dealing with time series. The following discussion will therefore focus on this specific data type, except when stated otherwise.

The most recurrent choice is the squared ℓ_2 norm (as in a least squares problem), which is also implicitly obtained by adopting the traditional likelihood model that defines observations as model predictions plus a Gaussian noise that is independent of the parameters. Yet any ℓ_p-norm, including the Euclidean distance ℓ_2, compares two data vectors element-wise. This represents a limitation when data points represent discretized signals, since they inherently exhibit a temporal structure that cannot be captured by simply comparing them at common values of the time coordinate. In fact

we will argue that ℓ_p norms ignore the dependence that exists between different points of the signals and can provide rather distorted distance measures in this setting.

Some literature exists about possible choices for distance measures between time signals for a variety of purposes such as pattern recognition, signal classification, detection, etc. [76]. Some still make use of the ℓ_2 distance, but only on slices of the signals, or after performing a circular shift of the time domain [50]. Other approaches propose counting the number of subsequences of the signals that are similar in an ℓ_2 sense [50]. Parameterizations of the signals (i.e., low-rank approximation) have also been proposed such that the comparison is made in this alternative domain rather in the original time or space one [11, 23, 75]. While attractive, these techniques are only relevant to specific applications and therefore tend to have limited applicability.

As mentioned earlier, what often results in discrepancies between modeled and collected data is some sort of warping along the time dimension. A broad set of literature exists around time warping, coming from all sciences and applications that need to deal with time series. The general theory around time warping is described in [106] in the broader context of functional data analysis. To warp a signal is to transform the input (e.g. time t) of a function $y(t)$, for instance with an invertible function h, to yield a warped signal $y(h(t))$:

$$t^* = h(t), \text{ and thus} \tag{3.1}$$

$$y^* = y(t^*) = y(h(t)) \tag{3.2}$$

where the h is chosen to satisfy a specific criterion. For instance, when used to build a misfit a function, h is chosen within some class $\mathcal{H}$ to *minimize* some notion of distance between *two* signals (e.g., model predictions and data). In formulae, a misfit $\text{dist}(\cdot, \cdot)$ between model predictions $u(t)$ and data $y(t)$ that allows for warping is:

$$\text{dist}(y(t), u(t)) = \arg\min_{h \in \mathcal{H}} C(y(h(t)), u(t)) \tag{3.3}$$

where is $\mathcal{C}$ is a cost function that the analyst chooses according to context and most

typically is some ℓ_p norm. The most basic type of time warping is the so-called shift registration. This kind of correction only affects the phase of each curve y, by shifting the respective support of a constant δ and using ℓ_2 as a cost function. More sophisticated choices for $\mathcal{H}$ and $\mathcal{C}$ are possible, such as those used in the well known framework of dynamic time warping (DTW) [95]. In a discrete setting, DTW allows for assignments of a point $y(t_j)$ to one or more points $u(t_i)$ of the comparing signal as long as the monotonicity of the mapping is not violated, i.e., $t_j \geq t_i$. While the cost function continues to be the ℓ_2 norm, efficient dynamic programming algorithms exist to tackle the computational problem. A number of variations around classical DTW have also been proposed, such as applying it only to subsets of the signal or introducing some weighting coefficients on the assignment choices. While the monotonicity condition is physically interpretable as the requirement to preserve "causality" in the assignment, it induces, especially when the number of discretization points between the two signals being compared is the same, splitting of the mass associated to a point i of a signal to several points j of the other signal. This sort of assignment is not particularly meaningful physically, as it is equivalent to concentrating/collapsing the signal rather than simply readjusting the time-scale. In the field of full waveform inversion, the use of DTW as a means to avoid or mitigate cycle skipping has been investigated in [82].

In the next subsection we will present an alternative to building time-warping-based misfit functions, using optimal transport distances.

3.1.2 The transport-Lagrangian distance: definition and algorithms

Recent advances in the domain of optimal transport and its many applications have lead a number of contributions in the field of signal analysis. Optimal transport allows the type of across-coordinate comparison of functional data that we seek, with some distinctive features compared to dynamic time warping. Optimal transport (OT) is, in general (cf. the Kantorovich problem), a way of finding a *coupling* of two probability

measures that minimizes a certain total transportation cost [140, 102]. In the very specific case of discrete/empirical probability measures with equal numbers of equally weighted points in their support, the OT problem reduces to an *assignment* problem [102] between the points in the support of each distribution. The transportation cost is often taken to be the distance or the squared distance between these points; the associated minimum total cost, over all possible assignments, is then the 1-Wasserstein distance or the 2-Wasserstein distance, respectively.

A distinctive feature of Wasserstein distances versus dynamic time warping is that causality is not ensured. This may seem a limitation in its application to signal comparison because of the inherent sequential nature of time signals. However, when dealing with model misspecification this aspect can actually be beneficial in that inconsistencies in the modeling can produce anticipation or delay in the reproduction of some parts of the observed signal.

One way of relating the OT problem to the comparison of time-dependent signals is to treat the signals as univariate probability density functions. For the resulting OT problem to even have a solution, however, it is necessary for these input signals to be normalized (i.e., integrating to one) and positive, as these conditions are necessarily satisfied by probability densities. Yet signals are not measures—i.e., they do not in general sum/integrate to one and are not in general non-negative. A common workaround to this problem is to shift the signal along the ordinate axis to make it positive and then divide it by the sum of all of its points [154, 89, 128]. Having the signal transformed in such a way also allows a fast, analytical, computation of the Wasserstein distance in 1-D. Attempts of using the Wasserstein distance in this fashion have been made in the field of waveform inversion too [154, 89]. Promising results were achieved in these works for velocity inversion. OT-based misfit functions have proven to be beneficial in terms of reducing cycle-skipping effects [20, 146]. While computationally convenient, the transformation of the signals that is required appears somewhat artificial and is not justified by the physics of the problem. In addition, the transformation can distort the signal, smoothing out amplitude versus frequency

differences [128]. In a general sense any *a priori* transformation of the data introduces the possibility of a number of artifacts in the results of the inversion that can be hard to predict and estimate. For this reason, when applied to field data these techniques may prove to be less reliable.

A different approach that avoids these pitfalls is to use the so-called transport–Lagrangian (TL) distance [129], which is a specific instantiation of the Wasserstein distance adapted to signals. Consider two real-valued signals $a, b : \mathbb{R} \to \mathbb{R}$. For simplicity, here we focus on the case where both signals have been discretized, the former on n points $\mathbf{t} = (t_i)_{i=1}^n$ and the latter on m points $\mathbf{s} = (s_j)_{j=1}^m$. Let $a(\mathbf{t}) = (a(t_i))_{i=1}^n$ and $b(\mathbf{s}) = (b(s_j))_{j=1}^m$. Then the TL distance can be written as the solution of the following minimization problem:

$$
\begin{aligned}
\mathrm{TL}_p^\lambda(a(\mathbf{t}), b(\mathbf{s})) &= \min_{P_{i,j}} \sum_{i,j} C_{i,j}^\lambda P_{i,j}; \\
\text{s.t.} \quad \sum_{j=1}^m P_{i,j} &= \frac{1}{n}; \\
\sum_{i=1}^n P_{i,j} &= \frac{1}{m}; \\
P_{i,j} &\geq 0;\ P \in \mathbb{R}^{n \times m}; \\
C_{i,j}^\lambda &= \lambda |t_i - s_j|^p + |a(t_i) - b(s_j)|^p;\ C \in \mathbb{R}^{n \times m}.
\end{aligned}
\tag{3.4}
$$

where C is a cost matrix; P a transport plan matrix; and $\lambda \in \mathbb{R}_{\geq 0}$ is a weighing parameter between the horizontal and vertical costs. This formulation can be interpreted in two different ways:

1. optimal transport between the graphs (2D) of a and b, i.e., $\{t_1 \times a(t_1), t_2 \times a(t_2), \ldots, a_n \times a(t_n)\}$;

2. optimal transport between two (1-D) uniform probability mass functions, with the cost defined as $C^\lambda(t_i, t_j) = \lambda |t_i - t_j|^p + |a(t_i) - b(t_j)|^p$.

This distance is particularly interesting as it avoids unnatural data transformations

while still allowing an OT formulation. In addition, while computing the Wasserstein distance in general discrete settings amounts to solving a linear programming problem (with $O(\max(n,m)^3)$ complexity, n and m being the dimensions of the discretized signals), for the special case of the TL distance with $n = m$, one can adopt more specialized algorithms that solve an assignment problem; in this case, the resulting optimal P are permutation matrices. Our algorithm of choice for such problems is the auction algorithm [13], which exhibits a nearly quadratic complexity or an average complexity of $O(n^2 \log(n))$ for problems with $n < 1000$ [114, 88]. In the rest of this thesis we will always consider $n = m$. Finally, the choice of the parameter λ is of crucial importance for a successful use of this distance. Generally speaking setting $\lambda \to \infty$ implies reverting to the ℓ_2 norm, while $\lambda \to 0$ allows for rather large amounts of horizontal transport, almost neglecting amplitude matching, which is, for most applications, the most informative feature of the data. Empirically we have found that a good choice for λ is that of ensuring the scale of the time vector values ($\mathcal{A}$) vs. that of the amplitude values ($\mathcal{T}$) are somewhat comparable i.e. $\lambda = \frac{\mathcal{A}}{\mathcal{T}}$. This is an accordance with related literature [88].

A rigorous discussion on the applicability of the TL distance as an objective function in deterministic seismic inversion has been conducted in [88]. Improvements in the convexity of the misfit have emerged as the primary effect of the choice of such a distance measure [154, 103].

3.2 A consistent Bayesian framework for optimal transport distances

In this section we intend to answer the following question: how can we build a coherent Bayesian framework around the TL distance as a misfit statistic? While maintaining the classic setup of additive Gaussian noise, we seek an alternative expression for the likelihood $p(\mathrm{TL}_2(\mathbf{y}, \mathbf{u})|\theta)$ (where $\mathbf{y}$ and $\mathbf{u}$ are the vectors containing the discretized form of the observed and model predicted signals, while θ are the model parameters).

Also note the choice of $p = 2$ to allow for more direct comparison with the ℓ_2 norm as a misfit statistic.

At this point, there are two main impediments that stand in the way of defining a coherent Bayesian framework for a TL-misfit, both in a well-specified and misspecified setting. First, calculating the TL-distance involves solving an optimal transport problem, which implies, in turn, a minimization problem: this non-linearity makes it difficult to derive an analytical expression for the likelihood $p(\mathrm{TL}_2(\mathbf{y}, \mathbf{u})|\theta)$. In the second place, we stated multiple times that it is of our interest to evaluate the robustness of such misfit measure in a misspecified context. From a rigorous standpoint however, the definition of likelihood assumes a context in which the data come exactly from the specified model. Therefore, even if we were able to obtain an exact expression for $p(\mathrm{TL}_2(\mathbf{y}, \mathbf{u})|\theta)$, this would not mean that the same expression could be used in a misspecified context without introducing some sort of inconsistency.

3.2.1 Gibbs posteriors

We therefore seek an alternative framework in which both misspecification and the newly introduced misfit measure can be integrated. In the statistical literature, a posterior distribution obtained through this framework is typically referred to as the Gibbs posterior. A full derivation is contained in [16], but we recall here a short summary. The central idea is to define a loss function $\mathcal{L}(\pi, \mathbf{y}; p)$ over our prior beliefs $\pi(\theta)$, observations $\mathbf{y}$ and space of probability measures p over θ. We then claim that a valid update of our beliefs based on available data is given by:

$$\hat{p} = \arg\min_{p} \mathcal{L}(\pi, \mathbf{y}; p). \tag{3.5}$$

This claim is justified by the argument that, in general, between two measures p_1 and p_2, one would naturally prefer the one that produces a lower value of the loss function, given the same data-set. The authors also choose a specific expression for the loss function that contains both of the fundamental ingredients of a Bayesian update, i.e.,

balance between prior information (Kullback-Leibler divergence between p and $\pi(\theta)$) and adherence to observed data under the form of expected loss:

$$\hat{p} = \arg\min_{p} \left(\int \ell(\theta, \mathbf{y}) p(d\theta) + \mathcal{D}_{\mathrm{KL}}(p, \pi(\theta)) \right). \tag{3.6}$$

The function $\ell(\theta, \mathbf{y})$ is a generic measure of model-data discrepancy (more discussion later). The authors show that the minimizer $\hat{p}$ takes the form:

$$\hat{p}(\theta) = \frac{\exp\{-\ell(\theta, \mathbf{y})\}\pi(\theta)}{\int \exp\{-\ell(\theta, \mathbf{y})\}\pi(d\theta)}. \tag{3.7}$$

This expression can justify a prior-to-posterior update through an exponential form given a generic loss function (or misfit measure) $\ell(\theta, \mathbf{y})$. While not a rigorously Bayesian update, it still captures the two main ingredients of Bayesian inference and provides a rigorous argument for using an exponential pseudo-likelihood. Additionally, we note that if it is known that the data arose from a given family of distributions (e.g., $p(\mathbf{y}|\theta)$), then equation (3.7) reverts exactly to Bayes formula, by taking $\ell(\theta, \mathbf{y}) = -\log(p(\mathbf{y}|\theta))$. This ensures the expression above constitutes a rational update with any misfit measure both in the well and misspecified context.

In our experiments, we adopted a specific expression for the Gibbs posterior as outlined in [94] (already experimented in a seismic inverse problem in [112]):

$$p(\mathbf{y}|\theta) = s^N \exp(-s\, \mathrm{TL}_p(\mathbf{y}, \mathbf{u}(\theta))). \tag{3.8}$$

where N is the number of observations while the parameter s acts as scaling factor.

The role of the s parameter This parameter plays no role in the data-generating process but it is necessary to ensure the values taken by the loss functions (in this case the TL-distance) are of the right order of magnitude to produce meaningful posterior distributions after being exponentiated. The scaling is therefore not an ad-hoc manipulation of the data to achieve more desirable results, bur rather a necessary adjustment to integrate any given loss function with a prior-to-posterior update that

is not derived explicitly from a physical model. This is reflected in the computational scheme used to calibrate the amount of scaling: s can be treated as a hyper-parameter and estimated through a hierarchical Bayesian framework. We associate to s a Gamma distribution as a conjugate prior, which allows a Gibbs update [94] in a Markov chain Monte Carlo (MCMC) algorithm that otherwise uses generic adaptive Metropolis [62] for θ updates. The choice of values for the shape and rate parameters of the Gamma prior is particularly critical to obtainment of a meaningful posterior. These values need to be picked in such a way that whatever loss function $\ell(\theta, \mathbf{y})$ is chosen to be used in the Gibbs posterior, it will scaled appropriately to avoid $\exp(-s \cdot \ell(\theta, \mathbf{y}))$ being numerically insensitive to different values of θ, making inference unfruitful. In the following section we will discuss the reasoning behind the choice of the Gamma prior for s through a numerical example.

3.2.2 Likelihood free inference

As a counterpart to a Bayesian framework that requires the definition of a likelihood, or a substitute for it, we outline a number of options for what is known as likelihood-free inference, an increasingly studied area. While we will not adopt any of these strategies for the main application and experiments in this thesis, but we will demonstrate their validity in a number of synthetic examples at the end of this chapter.

Approximate Bayesian computation (ABC) [84] is a common likelihood-free framework and is implemented as follows:

Algorithm 1 ABC—Approximate Bayesian Computation

1: *Initial value:* Propose initial estimate θ^* and define dataset $y_{1:n}$;

2: *Generate model samples:* Draw $z_{1:n}$ samples from $f(\theta_0)$;

3: *Calculate distance:* $\mathcal{D}(y_{1:n}||z_{1:n})$;

4: While $\mathcal{D}(y_{1:n}||z_{1:n}) > \epsilon$:

   1. *Resample* θ^**:* Propose another θ^* from a prior $p(\theta)$;

   2. *Generate model samples:* Draw n samples from $f(\theta^*)$;

   3. *Calculate distance:* $\mathcal{D}(y_{1:n}||z_{1:n})$;

5: *Accept:* $\theta_{accpeted} = \theta^*$;

6: Repeat the process K times, where K is the number of $\theta_{accepted}$ to characterize the uncertainty.

The $\mathcal{D}(\cdot, \cdot)$ is a distance or discrepancy measure between model and data chosen by the analyst, while ϵ is the admissible discrepancy up to which a sample θ^* can be accepted. This kind of estimation procedure possesses theoretical guarantees together with some common pitfalls, mainly concerning the choice of ϵ and how this affects the approximation of the true posterior, as well as its use in high dimensional parameter spaces. We refer to [124] for further discussion. In our context the main advantage of using such method is that it eliminates the need to characterize the likelihood function (or a surrogate for it) for a statistical model involving the TL distance.

3.3 Evaluating inference results: posterior scoring metrics and objectives

While Bayesian inference has become a widely used in many applications, it is still not entirely clear what constitutes a "good" posterior: how much uncertainty is the right amount of uncertainty? Should the true value of the parameter always be expected

to lie in high probability regions of the posterior (e.g., at the center of a Gaussian posterior)? A number of answers exists in literature and their content largely depends on the more fundamental question: "What do we want to use the posterior for?"

3.3.1 Continuous rank probability scores

Fairly well known in Bayesian inference are the so called *scoring rules* [52]. A score $S(G, H)$ is a measure of predictive accuracy of a forecaster G, established through an inference procedure, with respect to H, the "perfect" forecaster (e.g., true data distribution). A scoring rule is said to be proper if $S(H, H) = \min_G S(G, H)$. In other words, a scoring rule assigns the lowest score to the case where G equals the perfect forecaster. Considering continuous distributions with a density, a perfect forecaster H would be $H(y) = \delta_{y=y_{\text{true}}}$, while G can be any distribution $p(y)$ like a posterior distribution. Some examples of scoring rules are:

- *Brier score*(quadratic):

$$S(G, H) = \int_{-\infty}^{+\infty} (\delta_{y=y_{obs}}(y) - p(y))^2 dy; \tag{3.9}$$

- *Logarithmic score*:

$$S(G, H) = -\log p(y_{\text{true}}); \tag{3.10}$$

- *Continuous ranked probability scores* CRPS :

$$S(G, H) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{y} p(z)dz - \mathbb{1}_{y \leq y_{\text{true}}} \right)^2 dy. \tag{3.11}$$

  Forecasters are CDFs (cumulative distribution function) instead of PDFs (probability density function).

All of these rules assign a score zero to the case in which the probability assigned by $p(y)$ of observing the true data y_{true} is equal to 1. Among these kind of scores of particular interest is for us the case of the CRPS score. This score compares the CDFs

of the perfect and inference-built forecasters instead of their PDFs, which presents a number of advantages: since the CDF is a monotone increasing function, subtracting the perfect CDF (a step function set at y_{true}) to the inference built CDF provides at the same time a measure of how much bias and variability is contained in the posterior distribution. By bias we mean here how distant is most of the mass of the distribution $p(y)$ from y_{true} and by variability how "spread-out" the posterior distribution is. These features are relevant in a data-predictive context in which we want to reproduce data that is as close as possible to y_{true}. Figure 3-1 provides a visualization of the concepts behind the CRPS. In practice, the real value of y_{true} is unknown and thus the perfect

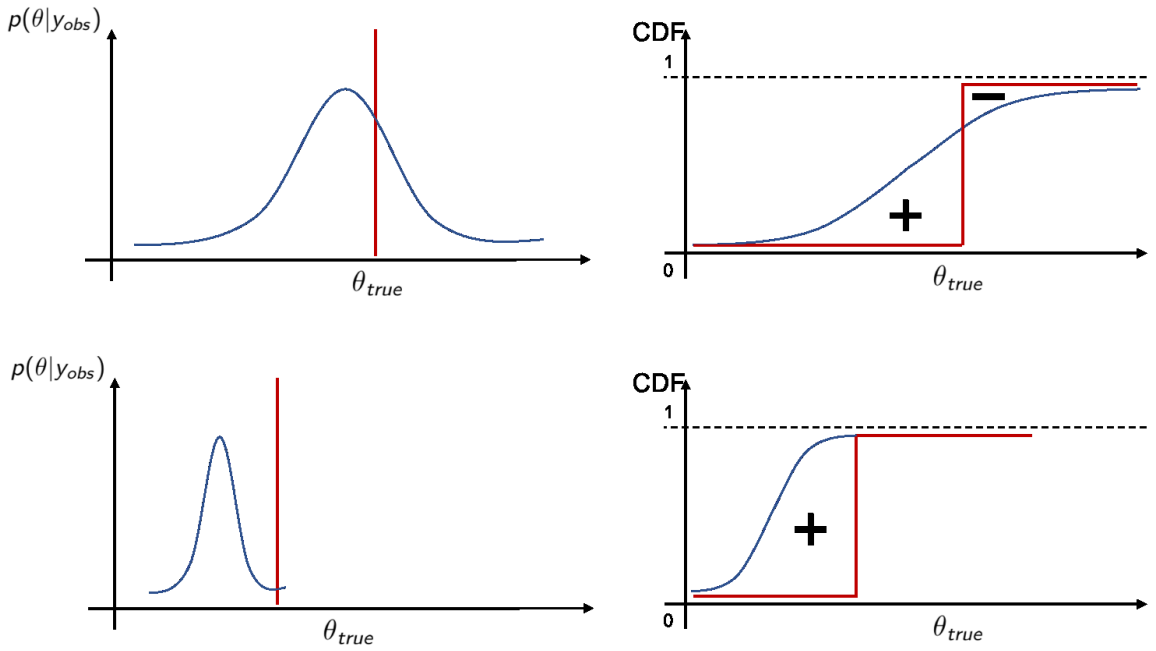


Figure 3-1: Bias (bottom) and variability (top) quantification in CRPS scores.

forecaster is approximated by building empirical distributions around "extra" or "newly collected" data. In the context of our experiment, instead, the CRPS scores would be of direct applicability since we actually know the true value of the quantity of interest $\mathbf{m}_{\text{true}}$ (not the data) and the trade-off between bias and variability of the posterior represents a valid way to compare distributions obtained through the two different misfit statistics.

3.3.2 Quantile rank statistics

Scoring rules tend to reward the predictive capability of a posterior, which is achieved with the least amount of bias and variability as possible. However, these properties may not completely overlap with other, statistically consistent, behaviors of posterior distributions. In particular a perfect forecaster may not exhibit what is known as the frequentist behavior of Bayesian credible intervals. In order to describe what this behavior is, we introduce another type of posterior-check: posterior quantile rank statistics [126] [27]. We stress that, contrary to the CRPS score, this is a *self-consistency* test that aims at answering the following question: is there any inherent bias in the way the posterior characterizes the uncertainty around the parameter space? Algorithm 2 describes the steps necessary to calculate quantile rank statistics and associated histograms for a specific test-case.

Algorithm 2 Quantile rank statistics

1: **for** $k \leq N_{rep}$ **do**
2: Draw $\theta^k_{\text{true}} \sim p(\theta)$
3: Generate data $\mathbf{y}^k \sim f(\mathbf{y}|\theta^k_{\text{true}})$
4: Estimate the posterior $p(\theta^k|\mathbf{y}^k)$
5: Draw M samples θ_i from the posterior distribution $p(\theta|\mathbf{y}^k)$
6: Calculate: $q_k = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\theta_i > \theta_{\text{true}}}$
7: $k \leftarrow k + 1$
8: **end for**
9: Plot histogram of $\{q_k\}$

The true values of θ_{true} should fall uniformly across the posterior credible set, just as, in a frequentist setting, an $\alpha-$confidence interval contains the true value $\alpha - \%$ of the times (Figure 3-2). This behavior translates into a uniform histogram over the sampled values of q: it indicates that the posterior distributions are neither overly biased towards one subset of the parameter set (Figure 3-3b), nor overly dispersive (Figure 3-3a), over-representing the amount of uncertainty in the problem.

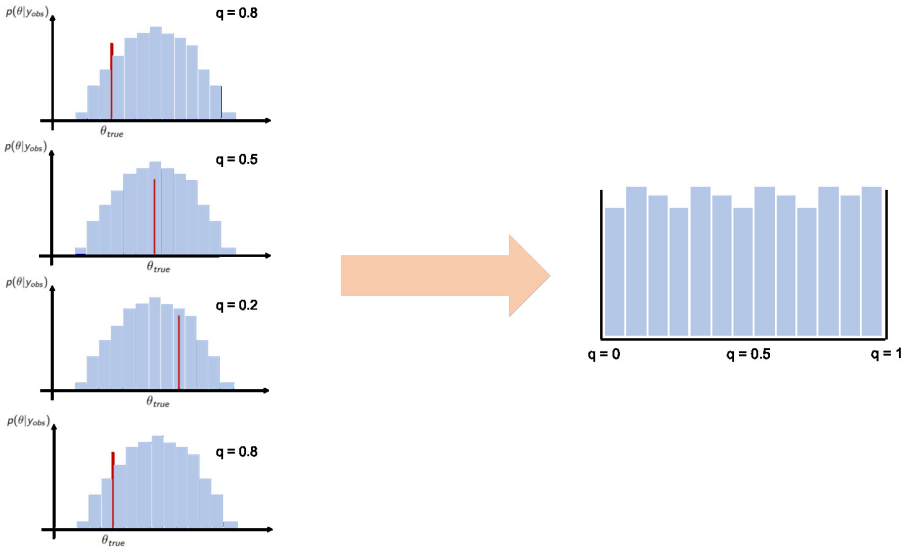
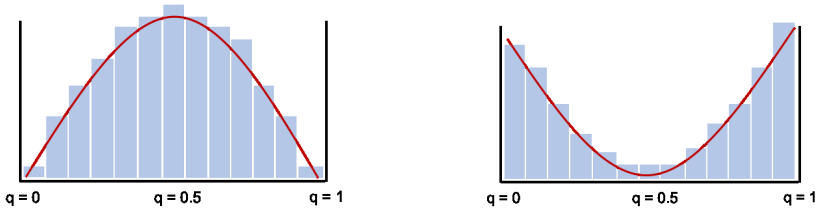


Figure 3-2: Quantile rank histogram building process under consistent conditions.



(a) Bell-shaped indicates an over-dispersed forecasting distribution.

(b) U-shaped indicates a biased forecasting distribution.

Figure 3-3: Non-uniform quantile rank-histogram shapes.

3.4 Numerical examples

In this section we present some numerical tests performed to start probing the validity and feasibility of the methodology exposed in the previous sections. We will start by describing a series of tests concerning the use of the TL distance in classification and inference problems. For all of the experiments below: let $\mathbf{t} = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n$ be a vector containing the time indices of the discretization. Let $\mathbf{y} \equiv y(\mathbf{t}) = (y(t_1), y(t_2), \ldots, y(t_n)) \in \mathbb{R}^n$ be the vector containing the values of the discretized signal used as data, and $\mathbf{u}$ the corresponding vector for the model).

3.4.1 Sine waves classification and inference with TL distance

Classification problems

We are first interested in testing the performance of the TL distance when it comes to classifying signals with respect to certain features. Throughout this entire section we deal with sine functions that present various differences in terms of phase, frequency and amplitude. The general expression of these functions we will refer to is:

$$\mathbf{u} = A\sin(\omega\mathbf{t} + \phi). \tag{3.12}$$

TEST A We have a reference signal $\mathbf{y_{ref}}$ of the form:

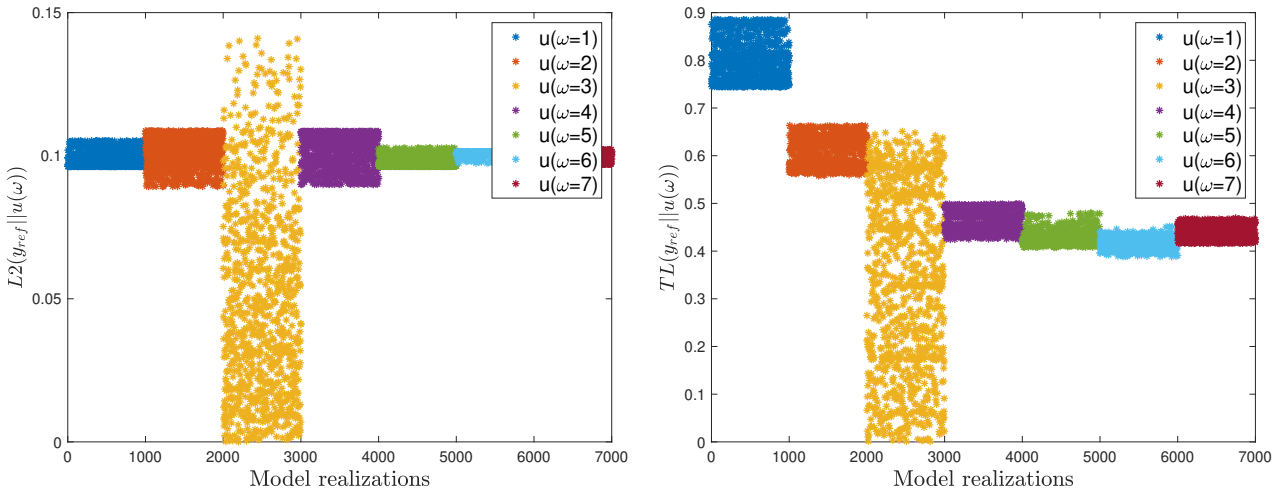
$$\mathbf{y}_{\text{ref}} = \sin(3\,\mathbf{t}). \tag{3.13}$$

We want to test how well the ℓ^2 and the TL distance allow for classification, with respect to $\mathbf{y_{ref}}$, of signals generated by the the model:

$$\mathbf{u} = \sin(\omega\,\mathbf{t} + \phi) \quad \text{where } \phi \sim \mathcal{N}(0,1) \text{ and } \omega = \{1,2,3,4,5,6,7\}. \tag{3.14}$$

The phase ϕ is drawn from a distribution to introduce some misspecification when comparing signals that might have the same frequency. In both cases the distance has been normalized by the number of discretization points. For the TL we use in this case $p = 2$ and $\lambda = 1$. 1000 samples have been drawn from the model for each value of ω. The distances are plotted in Figures 3-4a and 3-4b. Each point represents the distance (ℓ_2 or TL) between a realization of the model (3.14) and y_{ref}.

It appears clear how, on-average, the TL distance does a better job at distinguishing between the various frequencies given a random shift. When the frequency of the models is the same as the one of the reference signal ($\omega = 3$) the distances can reduce to almost zero, if the shift is not particularly significant. While the ℓ_2 distance can allow to detect that the true frequency is $\omega = 3$, it does not differentiate between the



(a) L2 distance vs model realizations

(b) TL distance vs model realizations

Figure 3-4: TEST A

other ω-s. The TL distance, assigns different distance values to different ω-s leaving the margin for even better detection if the the value of λ is carefully calibrated. The objective here is to show that the TL distance allows for better differentiation between signals given variations in a certain input parameter, not that it is the best tool detect the frequency of a signal per-se.

TEST B We have a reference signal $\mathbf{y_{ref}}$ of the form:

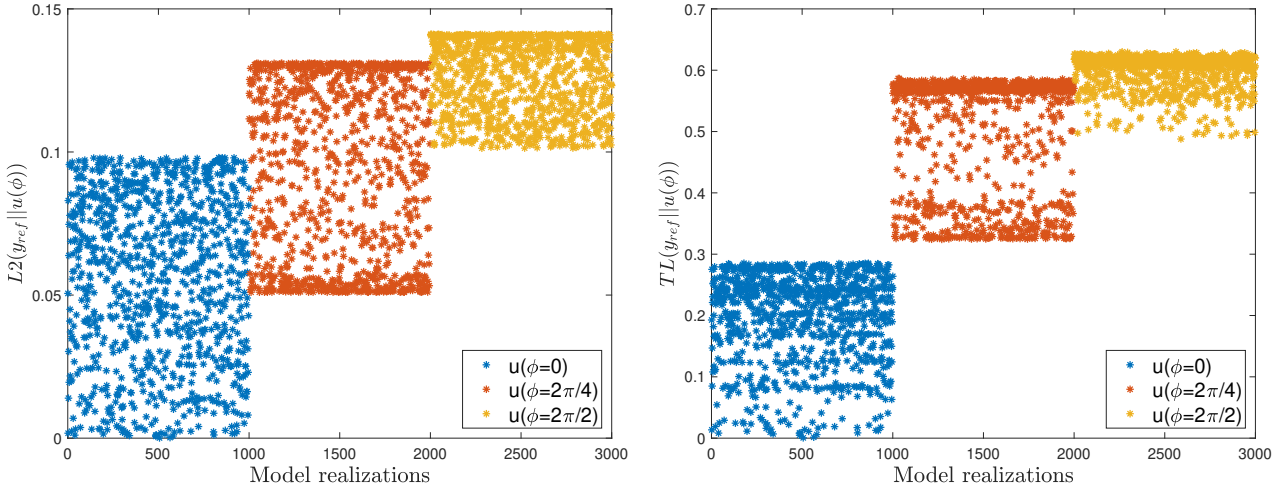
$$\mathbf{y}_{\text{ref}} = \sin(3\,\mathbf{t}). \tag{3.15}$$

We want to test how well the ℓ_2 and TL distances can classify signals generated by the model:

$$\mathbf{u} = \sin(\omega\,\mathbf{t} + \phi) \quad \text{where } \omega \sim \mathcal{U}(4.8, 7.2) \text{ and } \phi = \left\{0, \frac{2\pi}{3\omega}, \frac{2\pi}{2\omega}\right\}. \tag{3.16}$$

The classification is with respect to the delay ϕ, while the frequency is highly perturbed to verify how the distances perform in presence of some noise on ω. Again 1000 samples of $\mathbf{u}$ are drawn for each ϕ value. The results are reported in Figure 3-5a and 3-5b. Once again the TL distance seems to provide more separation between signals with

different delay values with respect to the ℓ_2.



(a) L2 distance vs model realizations

(b) TL distance vs model realizations

Figure 3-5: TEST B

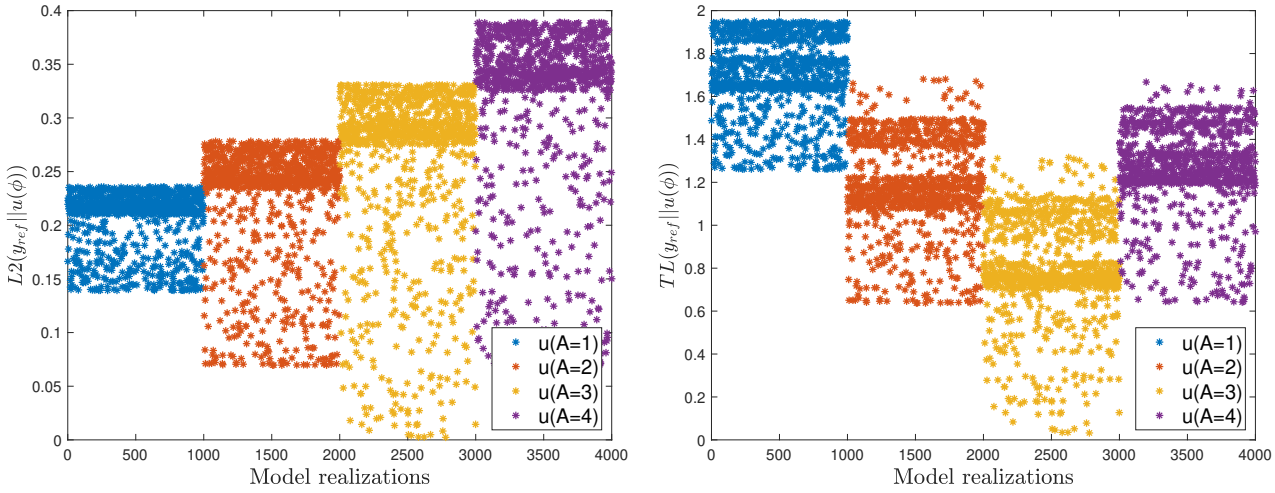
TEST C We have a reference signal y_{ref} of the form:

$$\mathbf{y}_{\text{ref}} = 3\sin(3\,\mathbf{t}). \tag{3.17}$$

We want to test how well the ℓ_2 and TL distances can classify, with respect to amplitude, signals generated by the model:

$$\mathbf{u} = A\sin(\omega\,\mathbf{t}) \quad \text{where } \omega \sim \mathcal{U}(1,3) \text{ and } A = \{1,2,3,4\}. \tag{3.18}$$

The frequency is perturbed to test the robustness of the distances to detect signals frequencies. The results are reported in Figures 3-6a and 3-6b. The TL distance performs dramatically better in distinguishing the amplitude of the signals regardless of the frequency perturbation. The ℓ_2 distance instead exhibits a higher degree of (relative) dispersion, especially when the amplitude increases.



(a) L2 distance vs model realizations

(b) TL distance vs model realizations

Figure 3-6: TEST C

Inference problem

We are interested in testing the TL based misfit function in an inference framework. To do so we have performed a number of tests, inspired by the classification problems in the previous section. Table 3.1 summarizes the test plan. There are 4 types of inference tests (i.e., different combinations of model and quantity of interest), for each of those types we tested 2 types of algorithms: Markov chain Monte Carlo Metropolis-Hastings (MCMC) with exponential likelihood [94] and approximate Bayesian computation (ABC). For each of these algorithms the classical ℓ_2 misfit measure is tested against the TL distance.

Before we proceed to the analysis of the results, we want to make explicit a number of technical details concerning the tests above:

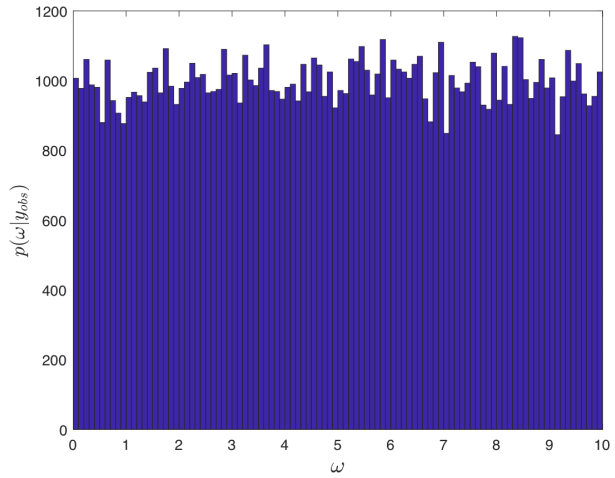
- *Computation of TL distance:* The formulation used for the TL distance is the one presented in equations (3.4) with $n = m$. The algorithm of our choice is the auction algorithm [13]. For the choice of λ, it is generally chosen to be around 1 in accordance to what explained in section 3.1.2.

- *Parameter prior distributions*: whenever a prior distribution needed to be defined, a proper uniform prior was adopted (details will be specified for each test case.

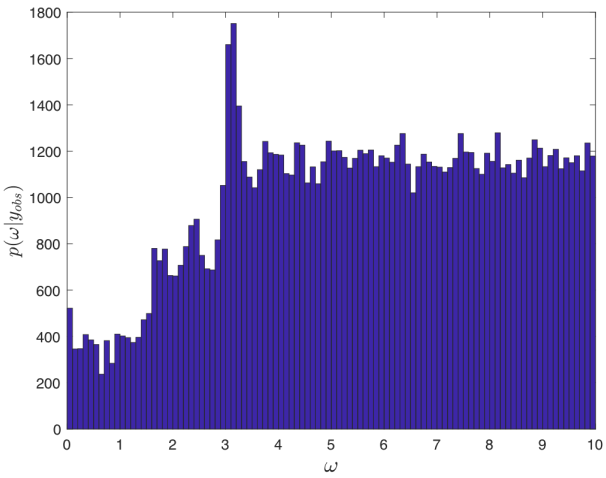
We can now proceed to discuss the results of the experiments.

TESTS A In this test the objective is to infer the frequency of an observed *sine* wave with a misspecified model: the model presents no phase shift, while the observed data has a value of $\phi = 0.7$ (chosen so that it is different from the period of the model *sine* function). The prior distribution of ω is uniform between 0 and 10. In Figure 3-7 we have reported the posterior and approximate posterior distributions for tests A.1, A.2, A.3 and A.4. The TL distance seems to outperform by far the ℓ_2 distance when used in the classical MCMC algorithm. When the ABC methods is used instead, we can see that the precision of the result highly depends on the choice of the acceptance threshold ϵ. For the ℓ_2 distance a very tight choice has produced a very narrow uniform distribution on the interval $[3.01, 3.23]$, which is close, although does not contain the true value. When the TL distance is used (A.2), we have that for a specific value of ϵ a number of peaks appears around specific values of ω. These partially reflect some of the results obtained in the classification exercise, although the true parameter value 3 does not emerge as clearly as in test A.1. For a smaller value of ϵ, however, a distribution similar to that obtained with the ℓ_2 distance can be achieved. While the TL distance seems to perform globally better than the ℓ_2, the ABC framework exhibits a certain amount of sensitivity relative to the choice of ϵ.

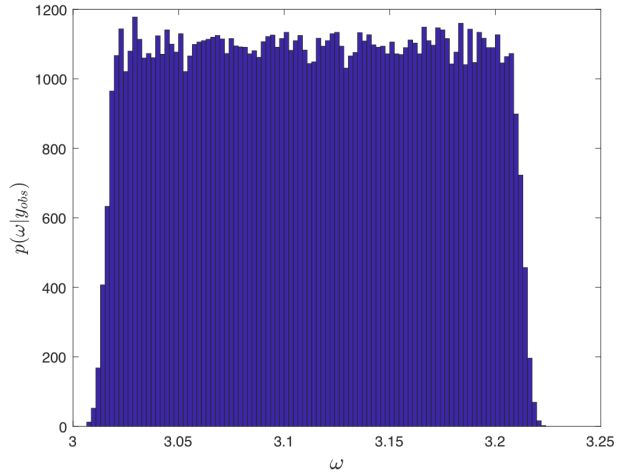
TESTS B In this test the objective is to infer the phase shift ϕ of an observed *sine* wave with a misspecified model: the model presents perturbations around the frequency value of the observed data, as specified in table 3.1. The prior distribution of ϕ is uniform between -2π and 2π. In Figure 3-8 we have reported the posterior and approximate posterior distributions for tests B.1, B.2, B.3 and B.4. The TL distance seems to outperform by far the ℓ_2 distance when used in the classical MCMC algorithm. When the ABC methods is used instead, we can see that the two distances perform equally well. While it is not particularly intuitive why in this case ABC performed better than MCMC, it may be worth mentioning that, when well calibrated, ABC can be more sensitive even to small distance differences, while MCMC does not



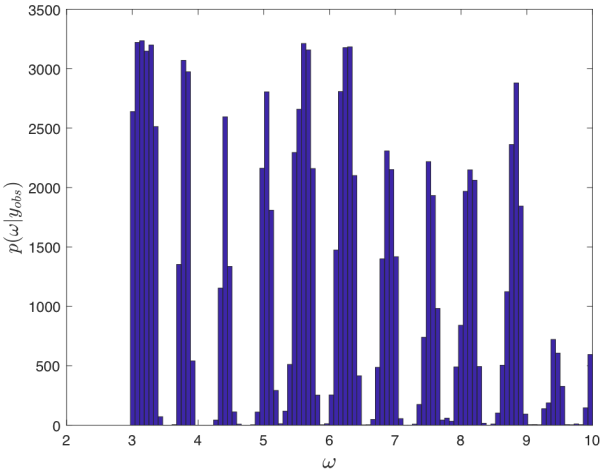
(a) TEST A.1: MCMC ℓ_2



(b) TEST A.2 : MCMC TL



(c) TEST A.5 : ABC ℓ_2

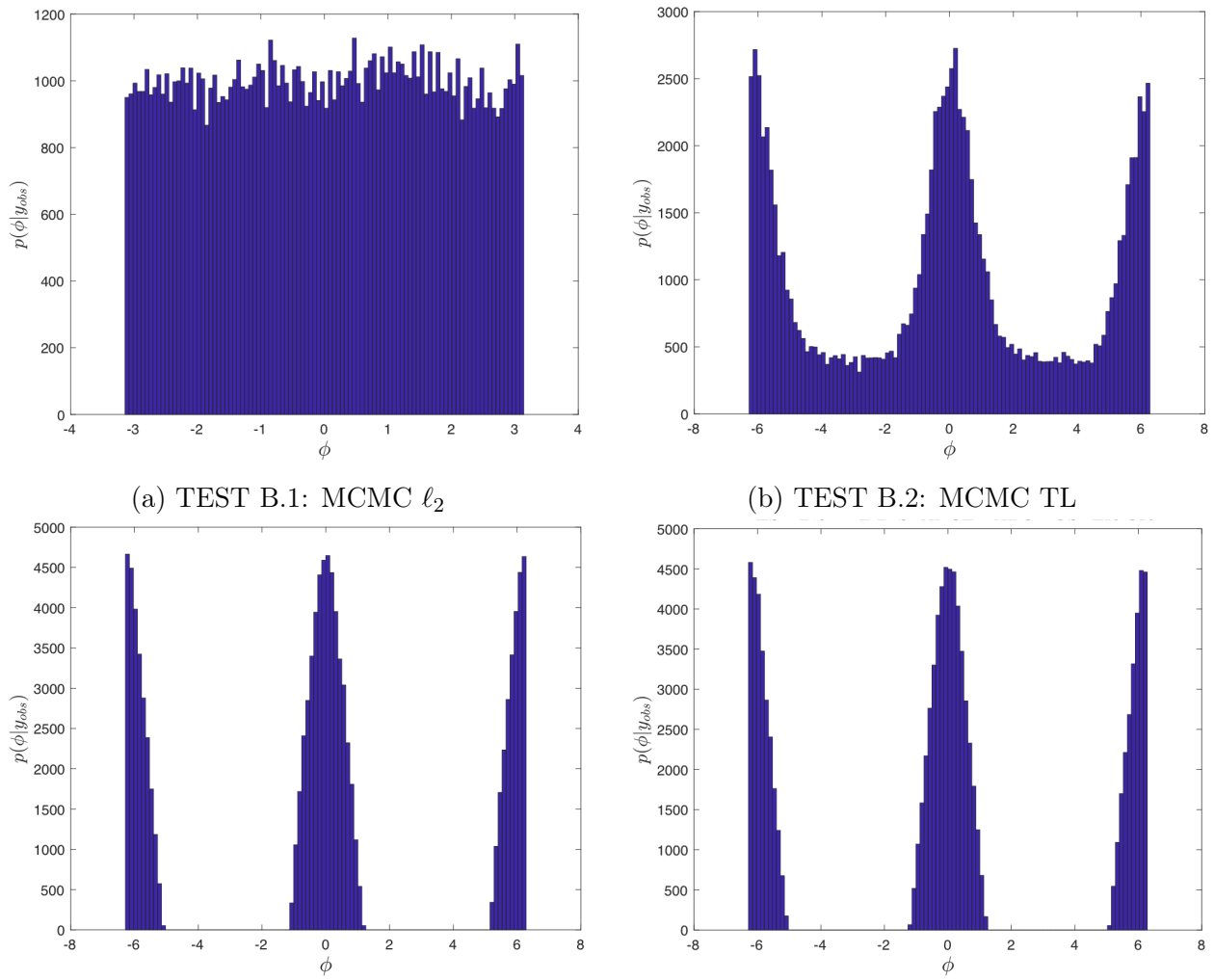


(d) TEST A.4 : ABC TL

Figure 3-7: TEST A - infer frequency, misspecified phase

operate based on a threshold-type mechanism.

TESTS C In this test the objective is to infer the amplitude A of an observed *sine* wave with a misspecified model: the model presents perturbations around the frequency value ω of the observed data, as specified in table 3.1. The prior distribution of A is uniform between 0 and 10. In Figure 3-9 we have reported the posterior and approximate posterior distributions for tests C.1, C.2, C.3 and C.4. The TL distance seems to outperform the ℓ_2 distance in all contexts.



(a) TEST B.1: MCMC ℓ_2

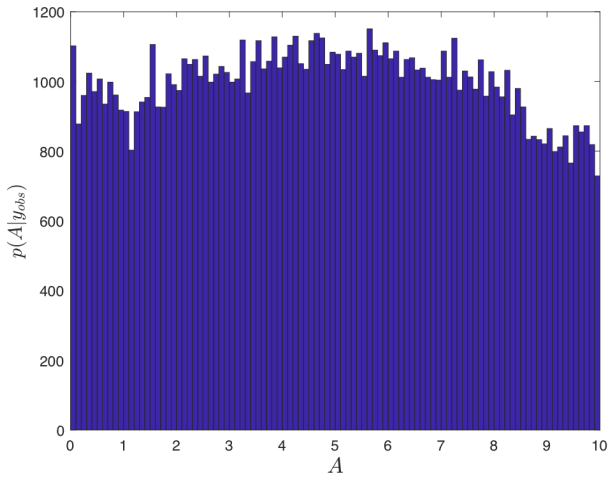
(b) TEST B.2: MCMC TL

(c) TEST B.3: ABC ℓ_2

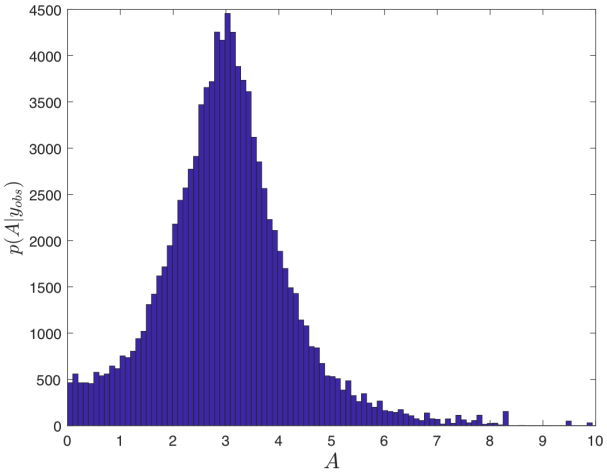
(d) TEST B.4: ABC TL

Figure 3-8: TEST B: infer phase, misspecified frequency

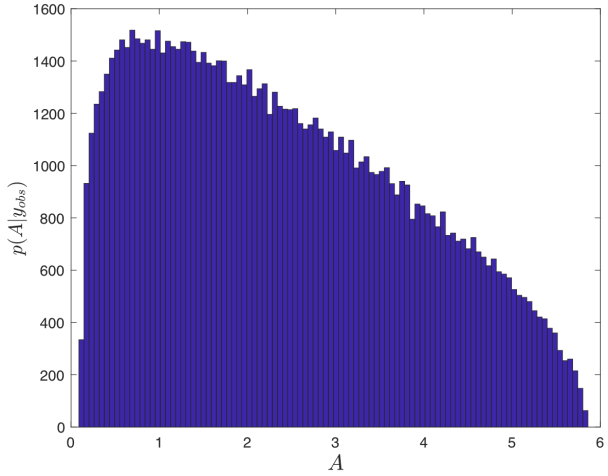
TESTS D In this test the objective is to infer the amplitude A of an observed *sine* wave with a misspecified model: the model presents perturbations around the phase shift value ϕ of the observed data, as specified in table 3.1. The prior distribution of A is again uniform between 0 and 10. In Figure 3-10 we have reported the posterior and approximate posterior distributions for tests D.1, D.2, D.3 and D.4. The TL distance seems to outperform the ℓ_2 distance in the classical Bayesian framework, while in the ABC context the results appear comparable for appropriate choices of ϵ.



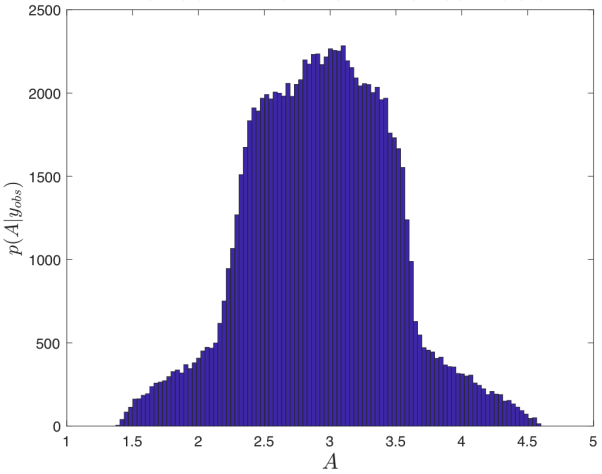
(a) TEST C.1: MCMC ℓ_2



(b) TEST C.2: MCMC TL



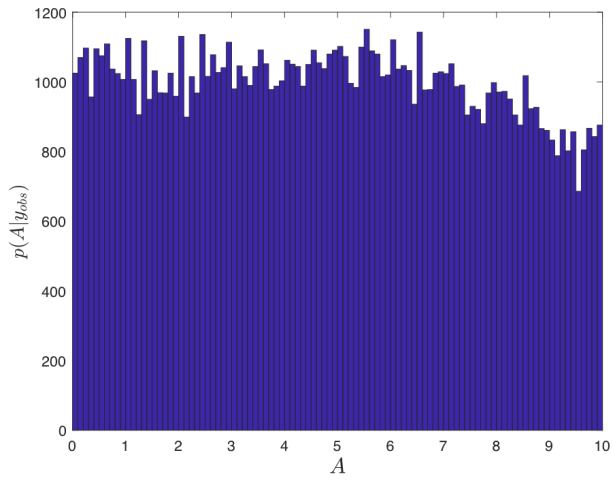
(c) TEST C.3: ABC ℓ_2



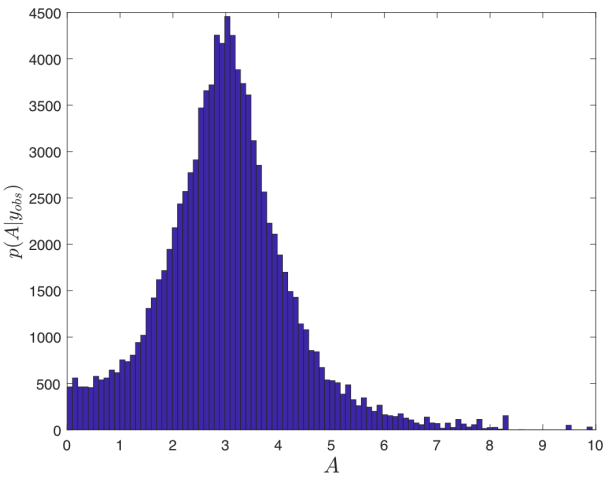
(d) TEST C.4: ABC TL

Figure 3-9: TEST C - infer amplitude, misspecified frequency

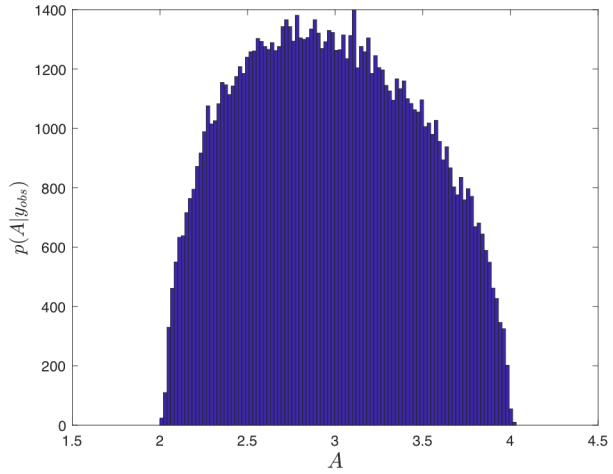
Conclusions: The TL distance seems to generally perform better than the ℓ_2 distance in a classical Bayesian inference framework. The ABC algorithm also presents satisfying results and avoids the problem of defining a likelihood function for the Wasserstein distance. However, it exhibits a certain degree of sensitivity to the choice of ϵ, whose choice only depends on computational power. Finally, it is important to note that in all test cases we were performing inference with a misspecified model: in this sense the TL distance seemed to be able to be more robust to model misfit, by providing a more suitable metric of comparison of the signals.



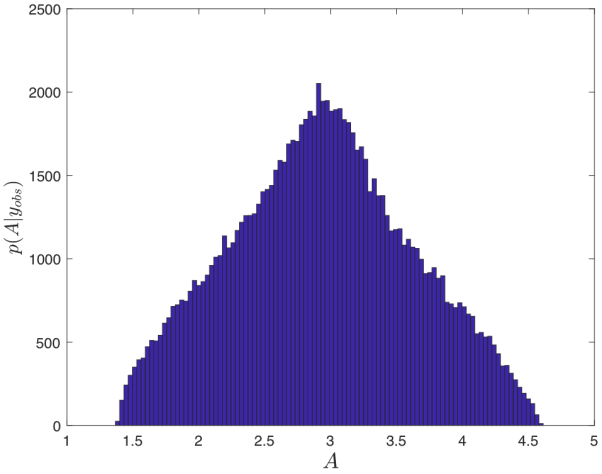
(a) TEST D.1: MCMC ℓ_2



(b) TEST D.2: MCMC TL



(c) TEST D.3: ABC ℓ_2



(d) TEST D.4: ABC TL

Figure 3-10: TEST D - infer amplitude, misspecified phase

3.4.2 Moment tensor inversion and seismic modeling with the reflectivity method

We are interested in evaluating the benefits of using the TL distance as a misfit statistic when solving the moment-tensor inverse problem in presence of model misspecification. To this purpose, we conduct an experiment using synthetically generated data from a layered-medium model. This model assumes that the waves travel through homogeneous elastic layers of different depth and velocity (one value for the velocity V_p of the primary waves and one value V_s for the velocity of the secondary waves).

The adopted solution to the forward problem is described in [17] and is a variation of a powerful and widely known method in seismology called "refelctivity method" [45]. In this application we assume to collect data from 4 different stations in each displacement direction (N-E-Z). These stations are located at $z = 0$ m from the Earth's surface and we can express their positions with respect to the epicenter of the earthquake in polar coordinates: station 1 - $r_1 = 5.6$ km, $\theta_1 = 60$ deg; station 2 - $r_2 = 3.5$ km, $\theta_2 = 110$ deg; station 3 - $r_3 = 4.1$ km, $\theta_3 = 250$ deg; station 4 - $r_4 = 5.3$ km, $\theta_4 = 280$ deg. The source is located at 1.1 km depth with duration 0.01 s. The entire waveform is recorded for 8.192 s, while we will only use the portion between 1 s and 7 s for the inversion. Model misspecification will be introduced by using a different velocity model for the data-generating process vs. the inference process.

We will now describe the velocity models that will be used throughout this thesis to create both well-specified and misspecified inference settings. In all of our experiments we will use a four-layered media velocity model (table 3.2) for the inference process. In a realistic setting, this model represents the analyst's best attempt at describing the geophysical characteristics of the terrain of interest. We call this model $\mathbf{V}_{4_{lay}}$. For the data-generating process we will instead use a model that exhibits 3 layers instead of 4, as specified in table 3.3. We call this model $\mathbf{V}_{3_{lay}}$. Note that the V_p/V_s ration has been kept constant across the models.

Note that the V_p/V_s ration has been kept constant across the models.

Nbr.	Thickness	V_p	V_s	ρ	Q_p	Q_s
	km	km/s	km/s	g/cm^3		
1	0.5	2.5	1.00	2.0	40	20
2	0.5	3.0	1.50	2.0	40	20
3	0.5	3.5	1.75	2.0	40	20
4	1.0	5.5	2.75	2.0	40	20

Table 3.2: Layer model used for inference.

Nbr.	Thickness	V_p	V_s	ρ	Q_p	Q_s
	km	km/s	km/s	g/cm^3		
1	0.8	2.5	1.00	2.0	40	20
2	1	3.2	1.60	2.0	40	20
3	0.7	5.5	2.75	2.0	40	20

Table 3.3: Layer model used for data generation

The objective is to test whether the TL distance performs better in terms of recovering the correct value of the moment tensor $\mathbf{m}_{\text{true}}$ compared to the the implicitly induced ℓ_2 norm of the additive Gaussian model (as described in 3.19).

Experiment 1 set-up

We first test the TL distance by examining its behavior in the misspecification setting as described in the experiment 1 prospect.

Experiment 1 Inference Procedure

1: Set $(\text{Strike}, \text{Dip}, \text{Rake}) = (300°, 20°, 150°) \rightarrow \mathbf{m}_{\text{true}} =$ [-0.50, 0.18, 0.32, 0.01, 0.74, -0.51];
2: Generate data $\mathbf{y}$ according to:

$$\mathbf{y} = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{3_{lay}}, \mathbf{t}) \cdot \mathbf{m}_{\text{true}}^T + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \tag{3.19}$$

3: Estimate the posterior $p(\mathbf{m}|\mathbf{y})$ assuming the following model for the data:

$$\mathbf{u} = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{4_{lay}}, \mathbf{t}) \cdot \mathbf{m}^T + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \tag{3.20}$$

The posterior will be a joint posterior over the 6 dimensional space of the moment tensor components (generally correlated). In order to evaluate the impact of the choice of the misfit statistic on the solution of the problem just described, we will integrate both the classic ℓ_2 misfit measure as well as the TL-based distance into the

pseudo-likelihood presented in the previous section. More explicitly:

$$\ell_2: \quad p(\mathbf{y}^k|\mathbf{m}^k) \propto \exp\left(-s\,||\mathbf{y}^k - \mathbf{u}^k||_2^2\right); \tag{3.21}$$

$$TL_2: \quad p(\mathbf{y}^k|\mathbf{m}^k) \propto \exp\left(-s\ \mathrm{TL}_2(\mathbf{y}^k, \mathbf{u}^k)\right); \tag{3.22}$$

The TL-based likelihood is the one derived in (3.8), where the parameter s acts as a scaling parameter. The ℓ_2-based likelihood is derived by simply substituting the ℓ_2 to the TL_2 misfit in the Gibbs posterior, although it is important to note that its analytical form corresponds exactly to the one that could be obtained by conditioning on the model parameters, starting from equations 3.19. In this case the s parameter could be directly interpreted as the model variance and would not need to be estimated through a hierarchical procedure (if assumed to be known). For consistency with the TL case, however, we treat it as a hyper-parameter and leave the discussion for the analytical solution of the linear Gaussian inverse problem for the end of this section. At the end of the experiment we will therefore have one posterior for each of the following cases:

$$\ell_2: \qquad p_{\ell_2}(\mathbf{y}^k|\mathbf{m}^k); \tag{3.23}$$

$$\mathrm{TL}_2: \qquad p_{\mathrm{TL}_2}(\mathbf{y}^k|\mathbf{m}^k). \tag{3.24}$$

Before discussing the results we briefly describe the settings for the actual algorithm used for Bayesian inversion.

Algorithm As already mentioned before, we implemented a Metropolis-within Gibbs scheme that updates $\mathbf{m}$ with a classic adaptive MCMC step, while for s it exploits the conjugacy of the Gamma prior. In particular we can sample s through a Gibbs update, meaning we can sample from the full conditional $p(s|\mathbf{m}, \mathbf{y}) = \mathrm{Gamma}(a, b + \ell(\mathbf{y}, \mathbf{u}))$. The term $\ell(\mathbf{y}, \mathbf{u})$ stands for whichever distance measure we are considering, either ℓ_2 or TL_2. The coefficients a and b are the shape and rate parameter of the Gamma prior on s.

Prior on moment tensor For the moment tensor prior we adopt a uniform distribution on the 6-dimensional ℓ_∞ unit ball (i.e., $\mathcal{U}(||\mathbf{m}||_\infty \leq 1)$).

Prior on s We set a Gamma prior on the scaling parameter s: Gamma(a, b). The rationale behind the choice for the shape (a) and rate (b) parameters is as follows: in order for $\exp\{-s \cdot \ell(\mathbf{y}, \mathbf{u})\}$ not to concentrate around 0 or $+\infty$ for any value of proposed $\mathbf{m}$, the monomial $s \cdot \ell(\mathbf{y}, \mathbf{u})$ needs to take values within the range $[-10, 10]$, at least for a subset of $||\mathbf{m}||_\infty \leq 1$. Depending on the average magnitude of the distance measure $\ell(\mathbf{y}, \mathbf{u})$, the Gamma prior must be chosen such that:

$$\mathcal{O}(s_{post}) \cdot \mathcal{O}(\ell(\mathbf{y}, \mathbf{u})) \approx \mathcal{O}(1) \tag{3.25}$$

where s_{post} is the s sampled from the conjugate posterior (i.e., $s_{post} \sim$ Gamma$(a, b + \ell(\mathbf{y}, \mathbf{u}))$); In our experiments, for both modes of misspecification we have that:

$$\mathcal{O}(\ell(\mathbf{y}, \mathbf{u})) \approx 10^{-2} \tag{3.26}$$

which in turn requires:

$$\mathcal{O}(s_{post}) \approx 10^{2} \tag{3.27}$$

Since $\mathbb{E}(s_{post}) = a \cdot (b + \ell(\mathbf{y}, \mathbf{u}))^{-1}$ and $\mathbb{V}(s_{post}) = a \cdot (b + \ell(\mathbf{y}, \mathbf{u}))^{-2}$, then an appropriate choice for the shape and scale parameter would be $a = 100, b = 1$. Given that $\mathcal{O}(\ell(\mathbf{y}, \mathbf{u})) \approx 10^{-2}$, this will result in:

$$\mathbb{E}(s_{post}) = 10^{2} \tag{3.28}$$

and approximately equal value for the variance.

Results

We want to compare two sets of 6 posterior distributions p_{ℓ_2}, p_{TL_2} and understand if and how the TL-based likelihood performed better than the ℓ_2 based one. At

this stage we can afford to visually inspect each of the posterior distributions and provide a qualitative judgement. In the following section however, we will broaden the experimental set-up and discuss ways of carrying out more systematic and quantitative evaluations of the quality of the posterior distributions.

In Figure 3-11 the marginals of p_{ℓ_2}, p_{TL_2} for each moment tensor component are shown side-by-side to facilitate comparison. It is quite clear how the TL-based posteriors seem to provide a better representation of the uncertainty around the true parameter values (red-lines). By "better" we mean in this case that TL-based posteriors are usually more centered around the truth and exhibit less spread around it. In contrast the ℓ_2-based posteriors are almost uniform for some parameters (m_{nn}, m_{zz}) or completely off-centered for others (m_{ez}, m_{ne}). These Figures however represent one

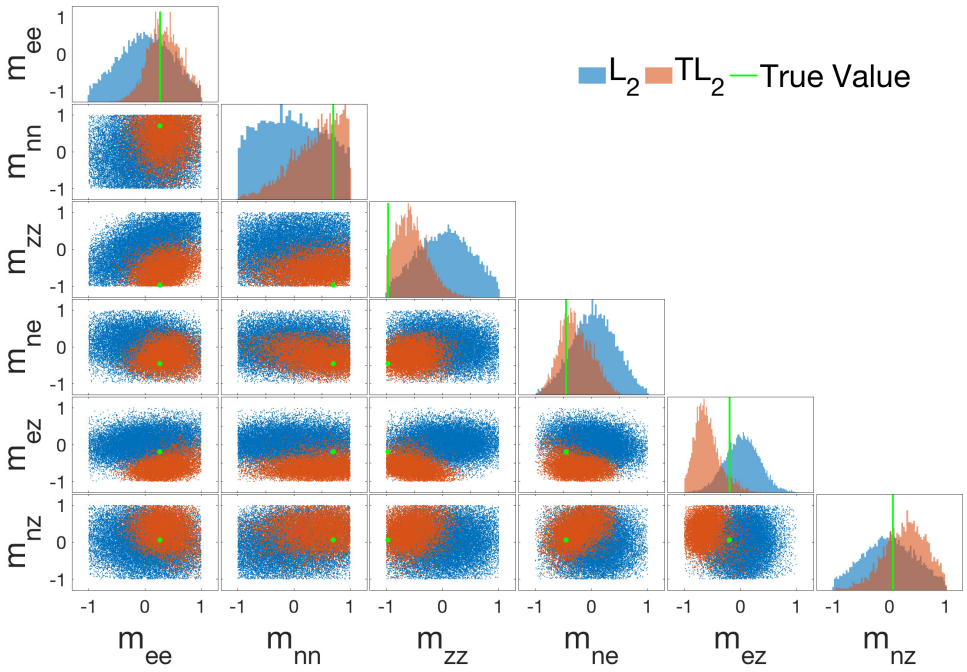


Figure 3-11: Sample p_{ℓ_2}, p_{TL_2} posteriors for misspecified model

specific instantiation of the problem and are therefore only anecdotal. In the following section a more systematic investigation of the behavior of the TL-distance compared to the ℓ_2 misfit will be conducted. In particular we will attempt to answer the more fundamental question of how to evaluate the quality of posterior distributions and,

more concretely, how to compare or rank them.

Experimental set-up extension for posterior scoring

We want to include CRPS scores in our experimental set-up to quantitatively assess the performance of the TL-misfit vs. the classic ℓ_2 distance under different instantiations of $\mathbf{m}_{\text{true}}$ for the velocity model configurations $\mathbf{V}_{4_{lay}}$ and $\mathbf{V}_{3_{lay}}$ defined in the previous section (experiment 2)

Experiment 2 CRPS scoring

1: **for** $k \leq N_{rep}$ **do**
2: Draw $\mathbf{m}^k_{\text{true}} \sim \mathcal{U}(||\mathbf{m}||_\infty \leq 1)$;
3: Generate data $\mathbf{y}^k$ according to:

$$\mathbf{y}^k = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{4_{lay}}, \mathbf{t}) \cdot \mathbf{m}^{k^T}_{\text{true}} + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}); \tag{3.29}$$

4: Estimate the posteriors $p_{\ell_2}(\mathbf{m}^k|\mathbf{y}^k)$ and $p_{\text{TL}_2}(\mathbf{m}^k|\mathbf{y}^k)$ assuming the following model for the data:

$$\mathbf{u}^k = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{3_{lay}}, \mathbf{t}) \cdot \mathbf{m}^T + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \tag{3.30}$$

5: Calculate the CRPS for the k-th posterior
6: **end for**

The results from this experiment can be analyzed in multiple ways, each revealing different pieces of information. First, for each of the posteriors obtained in experiment 2 we can calculate the CRPS score as follows:

$$\text{CRPS} = \frac{1}{N} \sum_i^N (F(\mathbf{m}_i|\mathbf{y}_{obs}) - \mathbb{1}_{\mathbf{m}_i > \mathbf{m}_{\text{true}}})^2, \tag{3.31}$$

where F is the empirical cumulative distribution function of a given posterior and the step function is the ideal CDF for the true value of moment tensor. As a first comparison measure we calculate the mean CRPS for each of the moment tensor

components obtained for both the ℓ_2 and TL-based posteriors.

$$\overline{\text{CRPS}} = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \text{CRPS}_k. \tag{3.32}$$

We report them in Figure 3-12 together with the associated estimator variance:

$$\sigma_{\overline{\text{CRPS}}} = \frac{1}{\sqrt{N_{rep}}} \sqrt{\sum_{k=1}^{N_{rep}} \frac{(\text{CRPS}_k - \overline{\text{CRPS}})^2}{N_{rep} - 1}}. \tag{3.33}$$

In order to make the comparison more significant we have repeated experiment 2 in a well-specified setting, i.e., with both the data and inference model Green's functions set to $\mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{4_{lay}}, \mathbf{t})$ and while using both the ℓ_2 and TL distance as misfit statistics. While in the well-specified setting both distances exhibit similar low scores, in the

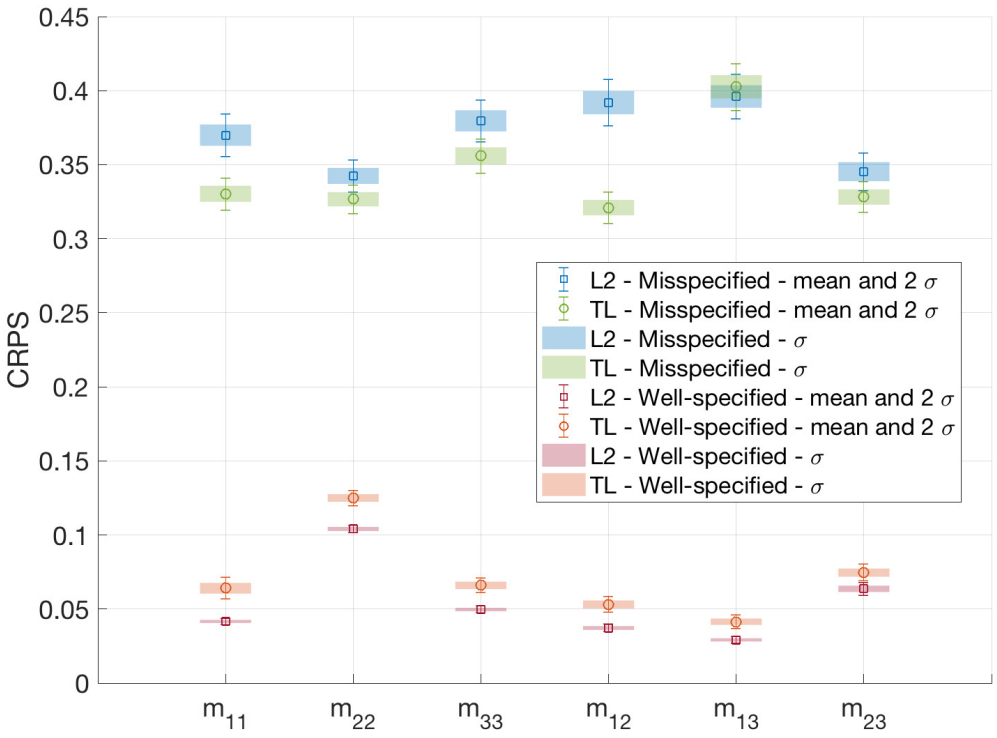


Figure 3-12: Mean CRPS scores in the well-specified (WS) and misspecified (MS) case and relative error bars.

misspecified settings the difference between the scores obtained with the ℓ_2 distance

and those obtained with the TL distance is quite significant. The distributions $p^1_{\ell_2}$ and $p^2_{\ell_2}$ achieve on average higher scores than the p^1_{TL} and p^2_{TL}, which indicates higher bias and/or variability. Average scores however do not provide a comprehensive image of the TL vs. ℓ_2 performance in terms of uncertainty quantification. Assuming that, as shown by the mean values, their behavior is almost identical in the well specified case, we focus on the misspecified setting. In this case, we are particularly interested in answering the following question: given the same $\mathbf{m}_{\text{true}}$ and the same velocity misspecification, how do the CRPS associated to the ℓ_2-based posterior compare to those obtained in the TL_2-based one? In particular, are the CRPS scores obtained for the $p_{\ell_2}(\mathbf{m}^k|\mathbf{y}^k)$ always higher than those obtained for the $p_{\text{TL}_2}(\mathbf{m}^k|\mathbf{y}^k)$? To provide a comprehensive and visual answer to this question we build the graph in Figure 3-13. For some randomly sampled pairs of CRPS, and every component of the moment tensor, we calculate the relative difference Δ_k and mid-point $\overline{\Delta}_k$:

$$\Delta_k = \text{CRPS}_k^{\ell_2} - \text{CRPS}_k^{\text{TL}_2} \tag{3.34}$$

$$\overline{\Delta}_k = \frac{\text{CRPS}_k^{\ell_2} + \text{CRPS}_k^{\text{TL}_2}}{2} \tag{3.35}$$

We then graph this information in the following fashion:

1. we select the moment tensor component of interest (horizontal axis);
2. if $\Delta_k \geq 0$, we plot a green box of height Δ_k with the centroid y-coordinate set at $\overline{\Delta}_k$. The width of the box is set to fixed value for graphical purposes only;
3. if $\Delta_k < 0$, we proceed as above, except that we will use the color red;
4. each box is filled with translucent color, which will produce darker shaded regions where multiple Δ_k will be centered around.

The characteristics of this plot allow for the following interpretation: the boxes being translucent, if in the majority of cases the difference in scores between the ℓ_2 and TL_2 posteriors is positive, then we will see a darker shade of green above the x-axis.

If instead the difference in scores is predominantly negative, the we will see darker shades of red. Positive Δ_k are predominant for all moment tensor components except m_{13}, showing a superior performance in terms of predictive capability of the TL-based posteriors. To achieve an even deeper analysis of the TL vs. ℓ_2 performance as a

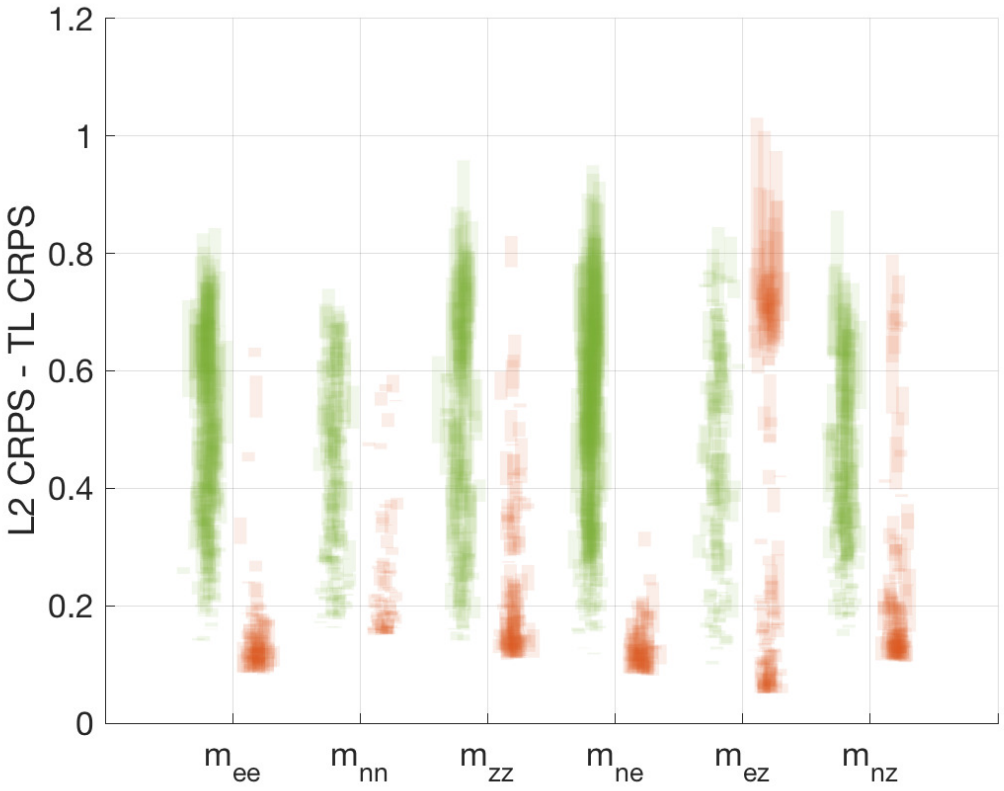


Figure 3-13: Box-plot for Δ_k for each moment tensor component in experiment 2. Red: TL score higher than ℓ_2, green vice-versa

misfit, we also plotted a histogram of the Δ_k per each moment tensor component (color-coded in the histogram) as well as a scatter plot of the Δ_k vs. $\bar{\Delta}_k$. While the histograms in Figure 3-14 clearly confirm the prevailing positive nature of the Δ_k already discussed in Figure 3-13, the scatter plots in Figure 3-15 provide additional information on the distribution of the Δ_k and the respective average score values $\bar{\Delta}_k$. One trend worth of observation is the fact that the differences in CRPS are much broader when positive, i.e., when the TL performance is superior, vs. when the ℓ_2 is performing better, in which case the difference in score is lower. We conclude by

tabulating the value of the estimated mean of Δ_k and relative standard deviation for each moment tensor component (Table 3.4).

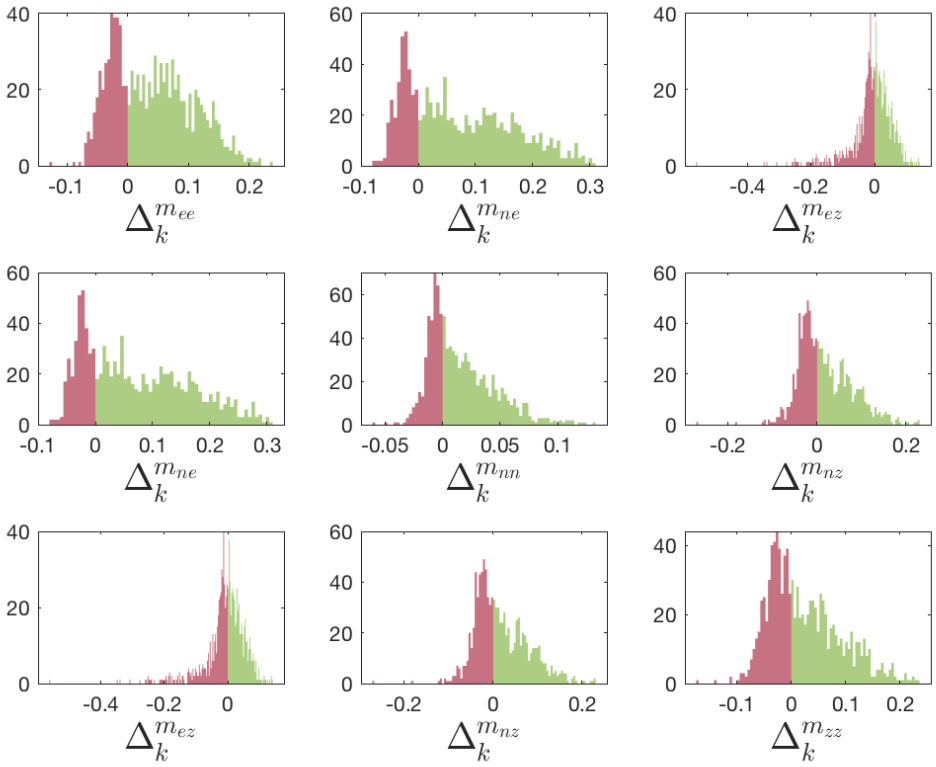


Figure 3-14: Histograms for Δ_k for each moment tensor component, arranged in a moment tensor matrix format for experiment 2.

Hierarchical model and analytical solution

In the previous section, we mentioned that in a well-specified setting it is possible to obtain an analytical solution to the linear-Gaussian inverse problem. In fact, assuming the noise level is known and fixed (i.e., $\Sigma = \sigma^2\mathbb{I}$) and an unbounded improper uniform prior, the posterior distribution is a truncated Gaussian with mean and variance as follows:

$$\mathbf{m}|\mathbf{y}, G(\mathbf{x}_{\text{true}}, \mathbf{V}_{\text{true}}) \sim \mathcal{N}((G^TG)^{-1}G^T\mathbf{y}, \sigma^2(G^TG)^{-1}). \tag{3.36}$$

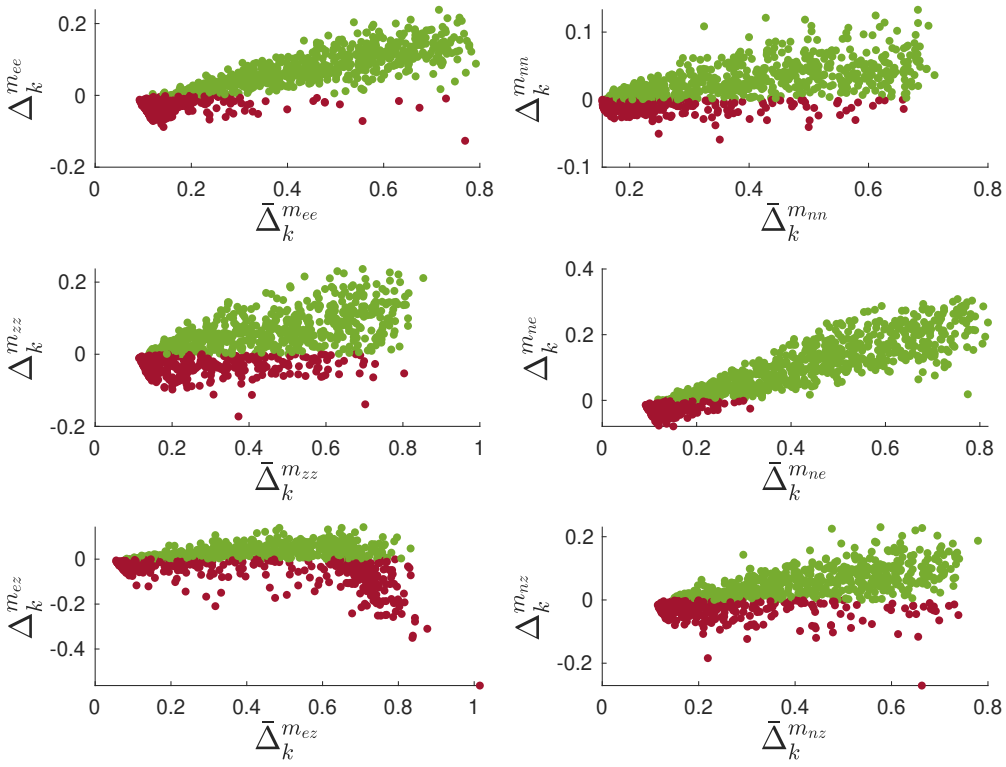


Figure 3-15: Scatter plot of Δ_k vs. y-coordinate set at $\overline{\Delta}_k$ for experiment 2.

Intuitively, this model should exhibit less variability with respect to the corresponding hierarchical one since the variance is known and there is also no need to add the scaling parameter s. The likelihood function inherently handles the scaling of the data-model misfit. We therefore repeat experiment 2 with the well specified velocity model and by using the analytic solution for the inverse problem. We then proceed with the calculation of the mean CRPS scores for this newly obtained set of posterior distributions and plot them against the ones coming from the hierarchical models (Figure 3-16). The analytical solution scores are expectedly much lower than the ones obtained in the misspecified case as well as those obtained in the well-specified case with a hierarchical model. They also exhibit much less variance. This result confirms the intuition that a less uncertain model produces better posteriors than a model that, from a theoretical standpoint, embodies more uncertainty given an additional parameter to be estimated. This behaviour can be further explained by referring to

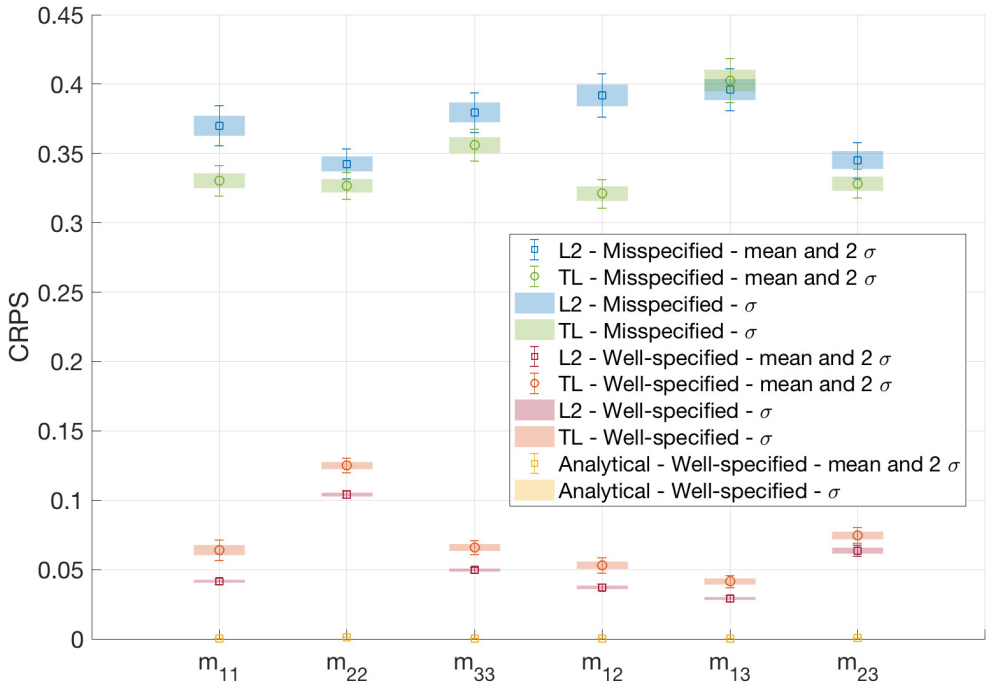


Figure 3-16: Mean CRPS scores in the well-specified (WS) and misspecified (MS) case and relative error bars, with the additional analytic solution to the WS case.

the scoring rules presented in section 3.3.2. Experiment 3 describes the steps necessary to calculate quantile rank statistics and associated histograms for our specific test-case.

We plotted the quantile rank statistics histogram for our experiment in the well specified case when using the ℓ_2 misfit both with the hierarchical model (Figure 3-18) and with the analytical solution (Figure 3-17). It can be observed that while with the analytical solution the histogram is uniform as expected, in the hierarchical model case it assumes a relatively narrow delta-shape around the center value 0.5. This result is consistent with the associated CRPSs: the analytical solution scores lower than the hierarchical model since it is behaves "perfectly" in Bayesian terms, i.e., exhibits good frequentist coverage over repeated realizations. Concretely, this means that while the true value does not always sit in the middle of the posterior distribution, the variance reduction spans several orders of magnitude compensating the increased bias. The hierarchical model posterior instead is more consistently centered around $\mathbf{m}_{\text{true}}$ at the price of over-dispersion, which induces a higher score. For completeness it is also

Experiment 3 Quantile rank statistics

1: **for** $k \leq N_{rep}$ **do**
2: Draw $\mathbf{m}^k_{\text{true}} \sim \mathcal{U}(||\mathbf{m}||_\infty \leq 1)$
3: Generate data $\mathbf{y}^k$ according to:

$$\mathbf{y}^k = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{\text{true}}, \mathbf{t}) \cdot \mathbf{m}^T_{\text{true}} + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \tag{3.37}$$

4: Estimate the posterior $p(\mathbf{m}^k|\mathbf{y}^k)$ assuming the following model for the data $\mathbf{u}^k = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}_{\text{true}}, \mathbf{t}) \cdot \mathbf{m}^T + \mathbf{e}$ where: $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$
5: Draw M samples $\mathbf{m}_i$ from the posterior distribution $p(\mathbf{m}^k|\mathbf{y}^k)$;
6: Calculate: $q_k = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}_{\mathbf{m}_i > \mathbf{m}_{\text{true}}}$
7: **end for**

interesting to check what the quantile rank histograms look like in the misspecified case as well. Figures 3-19 and 3-20 show the histograms in these cases. The results can be interpreted in the following way: under conditions of model misspecification the posterior distributions are, on average, biased and concentrate around the wrong values often enough for the histogram to assume the characteristic U-shape. This behavior is consistent with the Bernstein-Von Mises theorem discussed in section 2.1.1. While the histograms under these two cases look fairly similar it is worth nothing that the almost uniform histogram for component m_{22} in the ℓ_2 case is a byproduct of the fact that the associated posteriors are almost uniform. In fact, when a posterior is always uniform (bounded) and the true value is drawn from a uniform (bounded) prior, then the relative quantile rank histogram will also always be uniform. This may once again appear as a contradiction between quantile-rank checks that reward a totally uninformative posterior versus another kind of posterior checks (CRPS) that instead penalize the same posterior, since it is unusable from a forecasting point of view. Regardless of the specific moment tensor component, the CRPS clearly highlight a difference between the quality of the posterior distributions obtained with the ℓ_2 distance and the ones based on the TL_2 in a misspecified setting. However, the quantile rank histograms only slightly favor the use of optimal transport. This indicates that while the TL_2 can make inference more robust to misspecification in terms of predictive capabilities, it does not eliminate the misspecification itself.

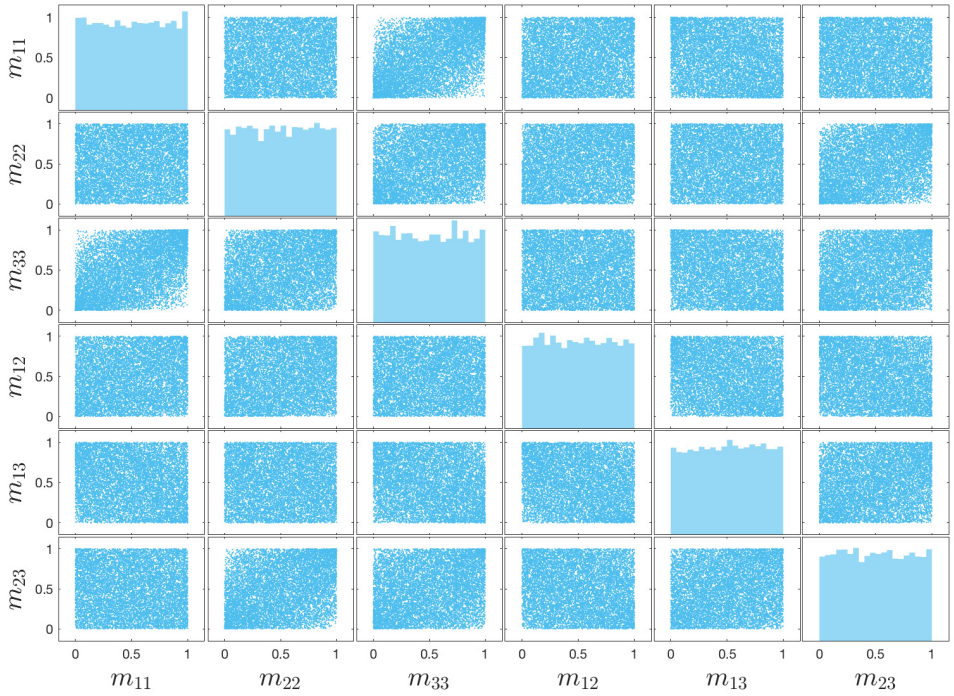


Figure 3-17: Quantile rank-histogram analytical model - well specified setting.

Trough this last exercise we have observed that that different posterior quality checks reward different behaviors of the posterior. In this regard, the rank-histograms are a more comprehensive measure of the "correctness" of an inference framework compared to the CRPS scores. However, as it often occurs in science and engineering, a "wrong" model may be more useful, under specific circumstances, than a theoretically "sound" one.

3.5 Conclusions

In this chapter we outlined the proposed methodology of this thesis for robust Bayesian inference based on optimal transport distances. In particular, we discussed the characteristics of the transport-Lagrangian distance, some algorithms to compute it, as well as the benefits it can bring to a specific category of misspecified inverse problems. We proposed and tested a number of statistically coherent frameworks for

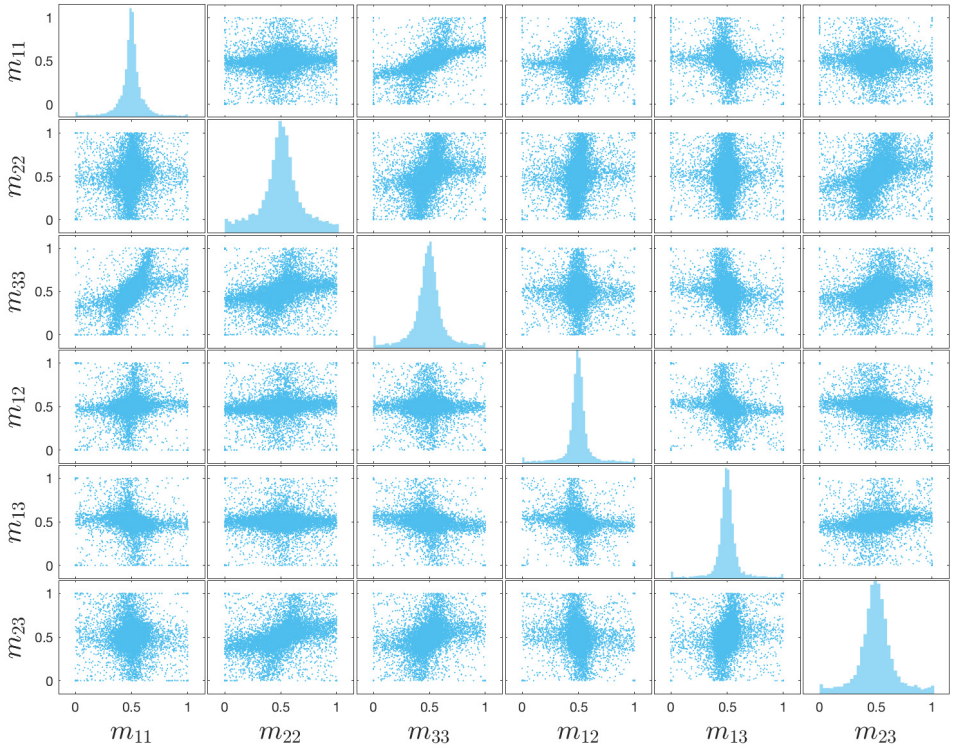


Figure 3-18: Quantile rank-histogram hierarchical model - well specified setting.

the integration of this distance in a Bayesian (or pseudo-Bayes) inference process. As an integral part of the framework, we also discussed some evaluation criteria to compare the statistical quality of the results obtained with the newly proposed method vs. those based on a classic ℓ_2-misfit. We also presented a proof-of-concept application of the method for moment tensor inversion with layered-media velocity models.

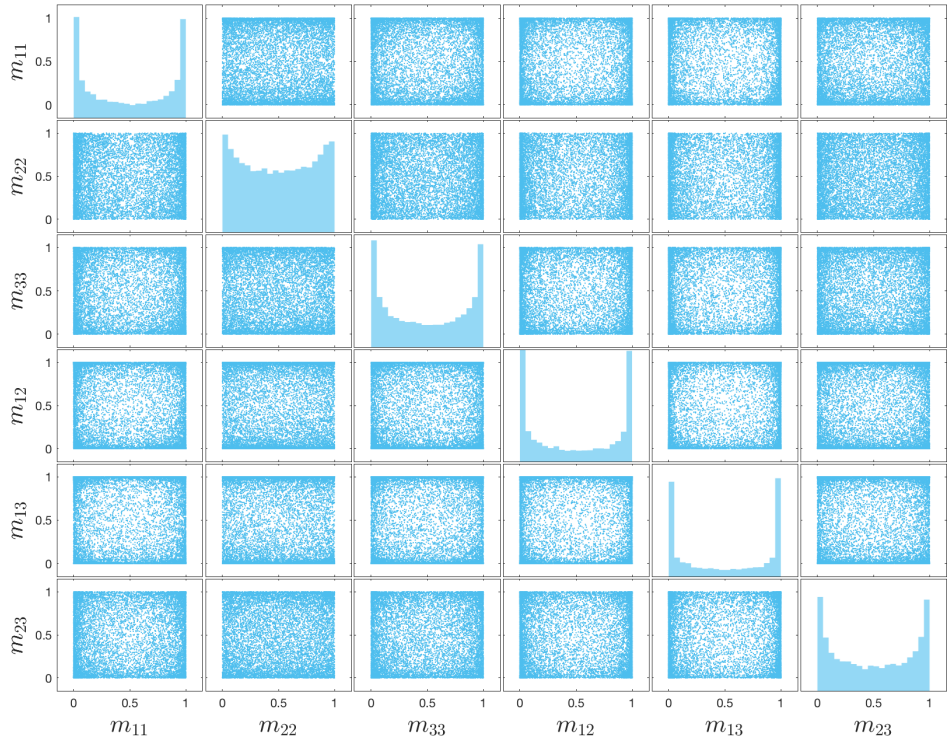


Figure 3-19: Quantile rank-histogram hierarchical model - misspecified setting ℓ_2.

MODEL	DATA	TEST Ref.	TEST QoI	ALGORITHM	DISTANCE
$\sin(\omega t + \phi) + e$ $e \sim \mathcal{N}(0, 0.1)$	$\sin(3t) + e$ $\phi = 0.7$	A.1	ω	MCMC	ℓ_2
		A.2		MCMC	TL
		A.3		ABC	ℓ_2
		A.4		ABC	TL
$\sin(\omega t + \phi) + e$ $e \sim \mathcal{N}(0, 0.1)$ $\omega \sim \mathcal{U}(1.9, 2.1)$	$\sin(2t) + e$	B.1	ϕ	MCMC	ℓ_2
		B.2		MCMC	TL
		B.3		ABC	ℓ_2
		B.4		ABC	TL
$A\sin(\omega t) + e$ $e \sim \mathcal{N}(0, 0.1)$ $\omega \sim \mathcal{U}(1, 3)$	$3\sin(2t) + e$	C.1	A	MCMC	ℓ_2
		C.2		MCMC	TL
		C.3		ABC	ℓ_2
		C.4		ABC	TL
$A\sin(2t + \phi) + e$ $e \sim \mathcal{N}(0, 0.1)$ $\phi \sim \mathcal{U}(-\pi, +\pi)$	$3\sin(2t) + e$	D.1	A	MCMC	ℓ_2
		D.2		MCMC	TL
		D.3		ABC	ℓ_2
		D.4		ABC	TL

Table 3.1: Inference tests with TL and ℓ_2 distance as misfit function

Parameter	m_{11}	m_{22}	m_{33}	m_{12}	m_{13}	m_{23}
Mean Δ_k	0.0396	0.0158	0.0237	0.0709	-0.0066	0.0171
Std.	0.0021	0.0009	0.0021	0.0029	0.0021	0.001

Table 3.4: Mean Δ_k values and associated estimator standard deviation - experiment 2.

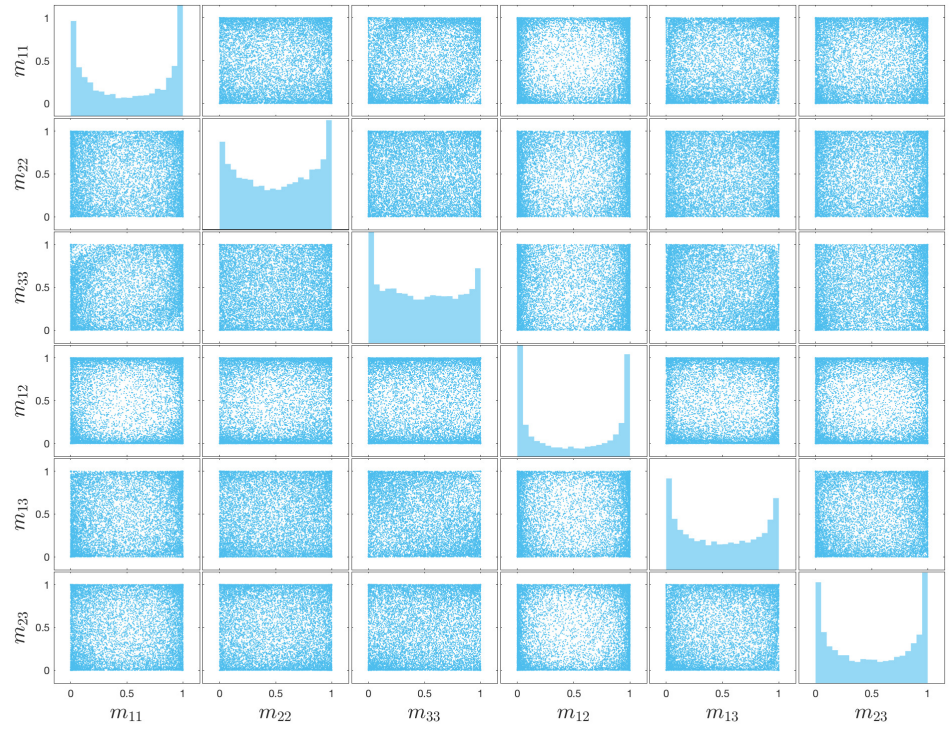


Figure 3-20: Quantile rank-histogram hierarchical model - misspecified setting TL_2 .

Chapter 4

An application: the SEG Overthrust model

Seismic inverse problems are a broad class of inverse problems that can be heavily impacted by model error. In this chapter we present a more realistic and complex example of velocity model misspecification in moment tensor inversion. For this specific inverse problem, a common approach is to generate model waveforms using a layered medium model (e.g., [34]). This model is often derived from well logs or from some other model of the subsurface, such as one derived from arrival-time tomography [59]. The model is an imperfect representation of the subsurface and cannot properly account for the 3D structure of a region. This adds considerable uncertainty to the results of moment tensor inversion. Looking at the effects of layered medium approximations to 3D velocity models is also, at present, an undeveloped area of research [133, 123]. This chapter will be articulated around the following topics:

- description of the velocity model setup and general inference scheme;
- discussion of numerical results;
- implications of the results on the recovery of non-double couple components in moment tensor inversion.

4.1 Velocity model and moment tensor setup

Our earthquakes are simulated with the 3D velocity model derived from the SEG/EAGE Overthrust Model [64, 7, 5, 6]. We choose this model because it contains structural complexity that, we assumed, cannot be easily represented using layered-medium models. We use a 15 by 15 km region located in the Southwestern portion of the Overthrust Model. The model extends to a depth of 4.7 km. Since only the P-wave velocity (V_p) model is given, we derive the S-wave velocity (V_s) using a variable V_p/V_s ratio in the range [2, 1.7], where V_p/V_s near the surface is close to 2 and it approaches 1.7 at the bottom of the model. The density model is obtained using the Gardner's relation ($\rho = 310V_p^{0.25}$).

Figure 4-1 shows the velocity at the source depth (1.1 km), the positions of the receivers (blue), which are located at the surface, and the source (yellow). Figure 4-2 shows East-West cross sections of the model taken at the source location, which is at the position of the yellow star. The source position was taken to be near the fault that cuts the anticline. We used a total of six stations located at the surface and surrounding the source. We simulated three-component waveforms for a single earthquake (strike, dip, rake = 40°, 50°, 280°, respectively) in the elastic 3D model using SPECFEM [74]. The source time history was taken to be a pulse that is nearly white between frequencies of approximately 1 and 13 Hz. These waveforms are taken as our earthquake waveforms. We derived layered-medium models to be representations of the 3D structure obtained from well logs. We took vertical profiles of the velocity and density models. We averaged the P-wave velocity over 500 m depth intervals to approximate how one might obtain a layered medium model from a well log. To this averaged (smoothed) model we added some noise equal to 2% at the top of the model and 10% at the base of the model to mimic increasing uncertainty in well logs with increasing depth. Further, we used a constant ratio of V_p/V_s of 1.73 to get the S-wave velocity. The density was taken to be constant at 2000 km/m3. We used vertical profiles at each station and the source location to yield a total of 7 layered velocity models. Figure 4-3 shows the source well-log

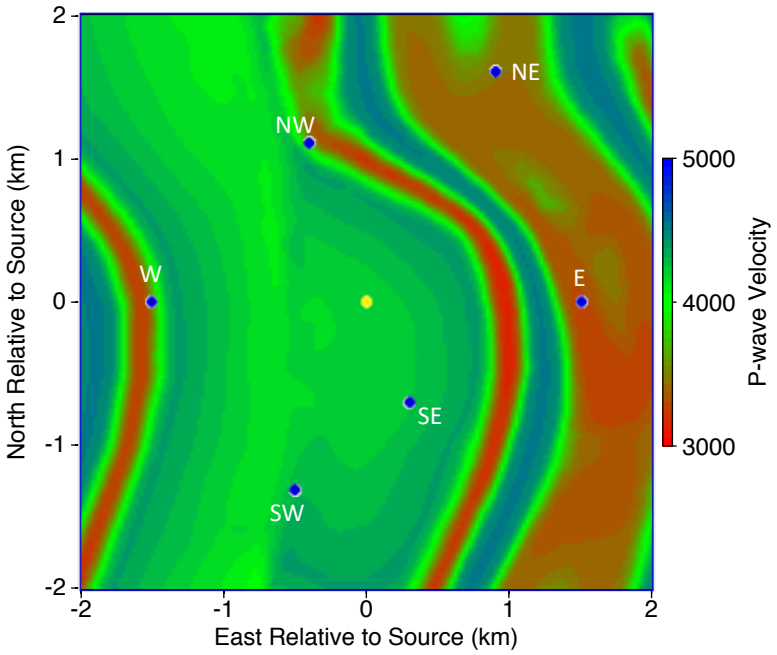


Figure 4-1: Horizontal cross section of the P-velocity model at the source depth (yellow dot). Locations of stations at the surface of the model are shown in blue.

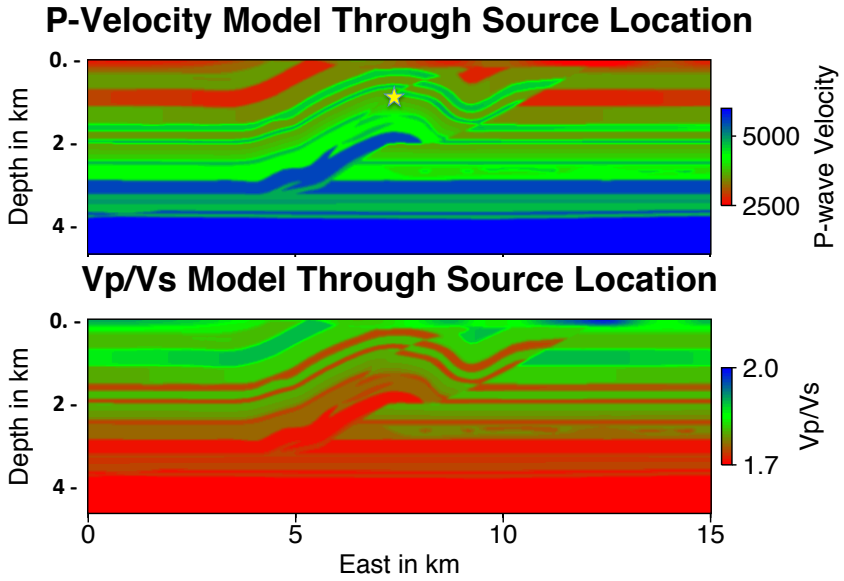


Figure 4-2: East-West vertical cross sections through SEG/EAGE Overthrust model at the position of the source (yellow star). Upper plot shows P-velocity model and lower plot shows ratio of P to S-wave velocities.

velocity profile (on the right) as well as the layered-medium models obtained by averaging model properties over depth (and adding some noise) at each of the other

well-logs locations. We simulated 3-component waveforms at each station for each of the seven velocity models using Axitra, a discrete-wavenumber reflectivity modeling approach [28]. We initially validated that waveforms simulated for an earthquake in a layered-medium model using both SPECFEM and Axitra were visually identical. Waveforms for each of the six moment tensor components at each station were then simulated using layered-medium models using the identical source time history as was used for the 3D earthquake simulation using SPECFEM. These waveforms were used for the inversion. Waveforms simulated using an identical layered medium model were used as moment-tensor Green's functions for each inversion test.

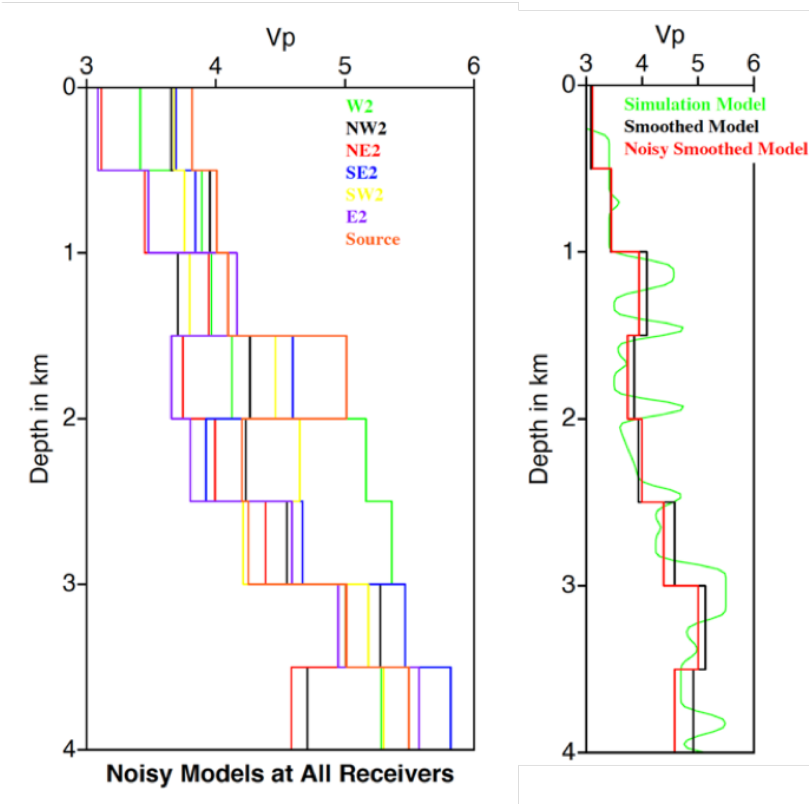


Figure 4-3: On the right: vertical velocity profile ("well log") of 3D model taken at source location (green) with smoothed (black) and noisy (red) smoothed profiles used to build the layered- media models. On the left: velocity profiles for layered medium models at each station location.

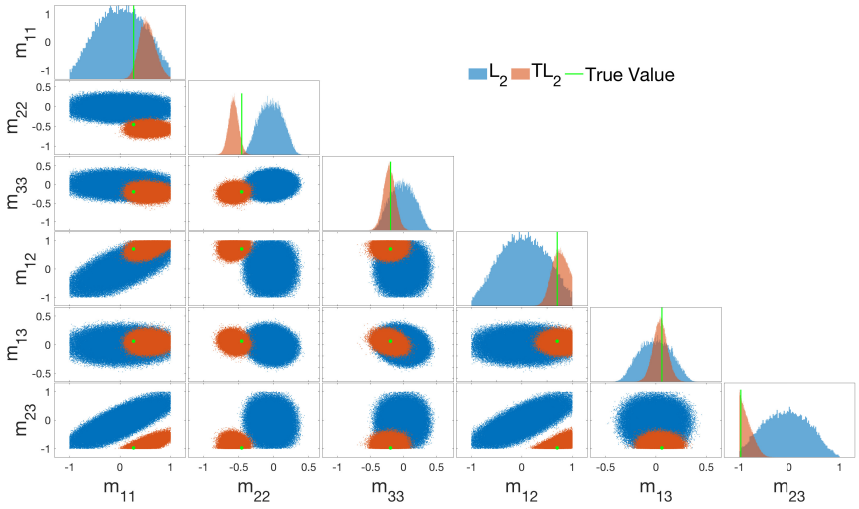


Figure 4-4: Source well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

4.2 Numerical results

Given the simulated data from the 3D model, we recover the moment tensor using each of the 7 layered velocity models, with either the TL or ℓ_2 distance. We illustrate the one and two-parameter marginal posteriors of $\mathbf{m}$ for the source velocity model in Figure 4-4. The TL approach provides significantly better recovery: it exhibits smaller variance and closer alignment with $\mathbf{m}_{\text{true}}$. We also report the combined TL-based and ℓ_2-based posteriors for the remaining velocity models in Figures 4-5, 4-6, 4-7, 4-8, 4-9, 4-10. For a more quantitative comparison, we report the average CRPS scores for each velocity model in Table 4.1.

For all of the alternative velocity models, the TL misfit provides better inference and uncertainty quantification for the moment tensor. The lower CRPS scores indicate that the TL-based posterior distributions are on average less biased, and exhibit less variance, than those obtained with the standard ℓ_2 distance. This translates into more reliable moment tensor estimates even in the present misspecified setting - i.e., when a realistic 3D velocity model is represented (incorrectly) by a layered medium model constructed from well logs.

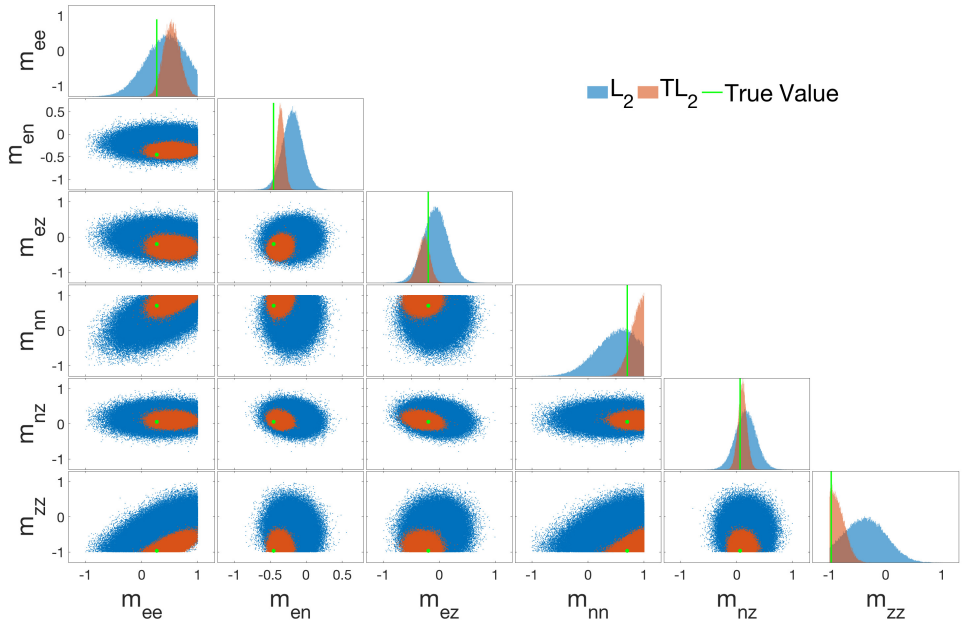


Figure 4-5: East (E) well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

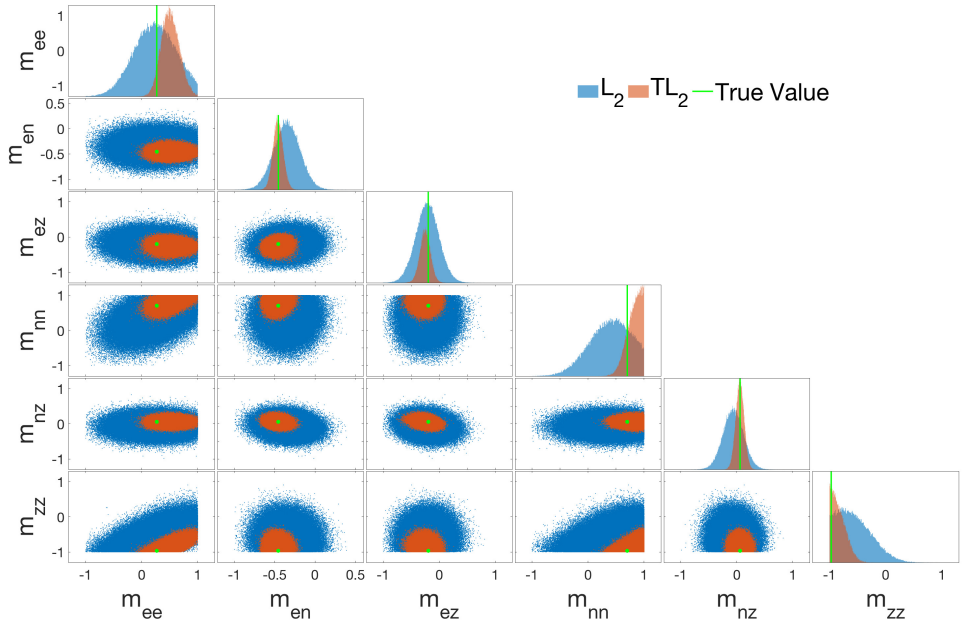


Figure 4-6: West (W) well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

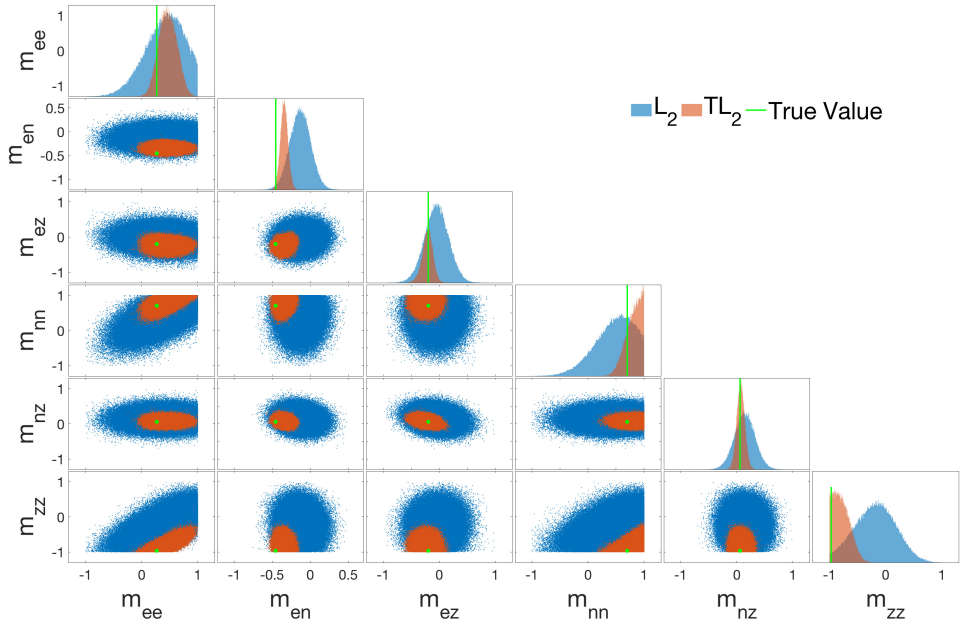


Figure 4-7: North East (NE) well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

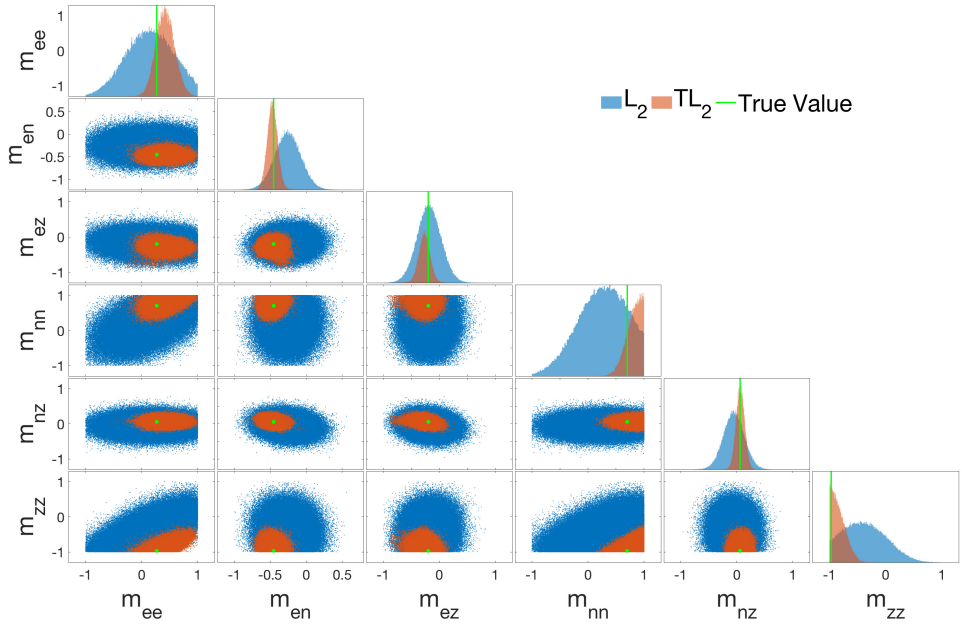


Figure 4-8: North West (NW) well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

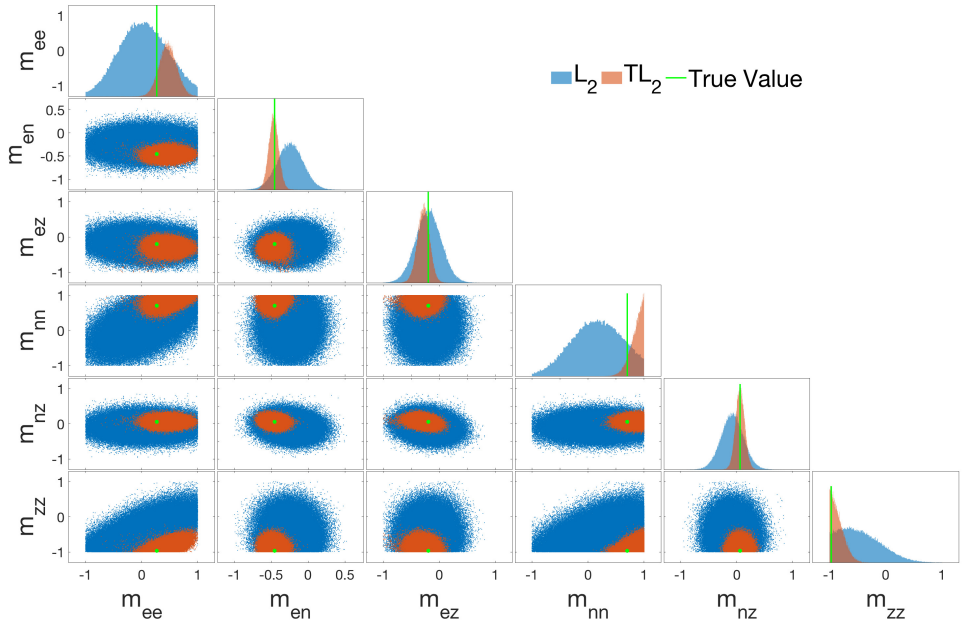


Figure 4-9: South East (SE) well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

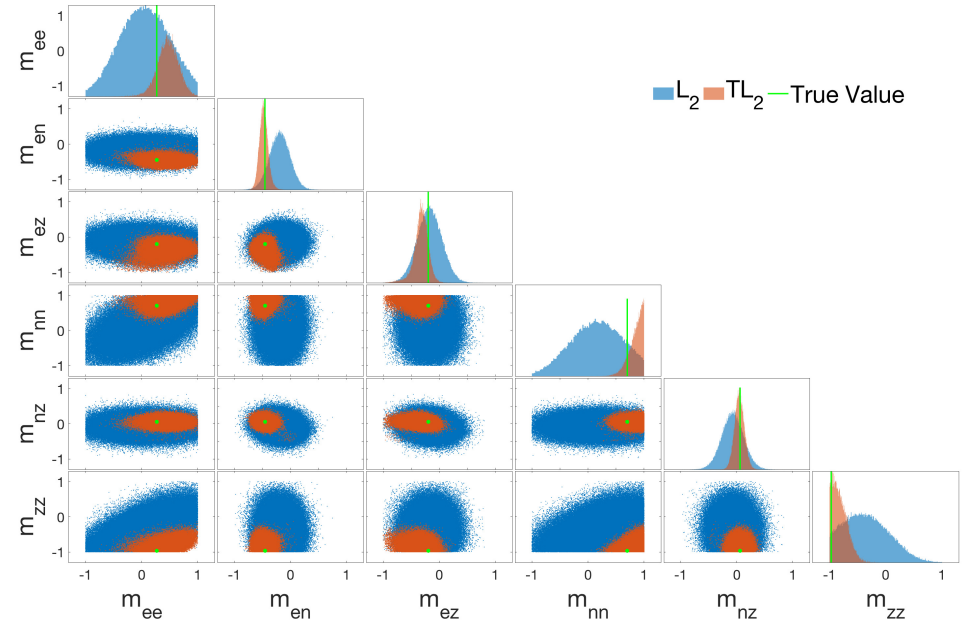


Figure 4-10: South West (SW) well log velocity model: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

Well log	**TL$_2$**	**ℓ_2**	**Δ_{ℓ_2-TL}**
NW	0.0528	0.1685	0.1157
NE	0.0637	0.1785	0.1317
SW	0.0710	0.2005	0.1295
SE	0.0637	0.1785	0.1148
W	0.0635	0.1173	0.0539
E	0.0837	0.1665	0.0828
SOURCE	0.0799	0.3008	0.2209
Mean (Std)	**0.0694** 0.0105	**0.1907** 0.0562	**0.1213** 0.0198

Table 4.1: Average CRPS scores for 1D marginal posteriors

By looking at the variability of the scores across the models, it also appears that the posteriors obtained with the ℓ_2 distance are a bit more sensitive to the velocity model used for inversion than the TL ones (CRPS standard deviation of 0.005 of the ℓ_2 vs. 0.001 for the TL). The interpretation of this behavior in geophysical terms requires further investigation, but indicates that TL distance is less sensitive than ℓ_2 to variations of the velocity model. This, in turn, suggests that OT misfit measures exhibit some robustness to variations in experimental design (i.e., choice of station and well log location).

It is also interesting to calculate the CRPS averages for each moment tensor element, averaging across velocity models. We report the results in table 4.2 It appears

m	**TL$_2$**	**ℓ_2**	**Δ_{ℓ_2-TL}**
m_{ee}	0.1410	0.1294	-0.0116
m_{en}	0.0399	0.1809	0.1410
m_{ez}	0.0455	0.0625	0.0169
m_{nn}	0.0951	0.2607	0.1656
m_{nz}	0.0202	0.0724	0.0522
m_{zz}	0.0744	0.4381	0.3637

Table 4.2: Average CRPS per moment tensor across velocity models.

that on average the m_{ee} posteriors obtained with the ℓ_2 exhibit lower CRPS scores than the TL-based ones. By simple visual inspection, it does appear that the ℓ_2 posteriors, while in general more dispersive (i.e. higher variance), exhibit proportionally less bias, which in turn produces a lower score. At the moment we do not have an explanation

in geophysical terms for this behavior. However, since it does not seem to be linked to any specific velocity model, possible causes may be related to the chosen network configuration.

Six-dimensional quantitative measures As an additional measure of result quality, we calculate the inner product between the true $\mathbf{m}_{tens}$ and each sampled $\mathbf{m}_{tens}$. This measure is of particular interest since it looks at the six components of the moment tensor jointly, rather than separately as the CRPS does. In fact a sample from the posterior has a geophysical meaning only when analyzed in its entirety. We report in table 4.3 the results for each velocity model. Through this 6-dimensional measure, the TL manifests itself as the clear winner, with TL-posterior samples scoring an average of 0.9528 vs. 0.6146. As an additional measure of closeness of the posterior samples

Well log	$\mathbf{TL}_2$	$\boldsymbol{\ell_2}$	$\boldsymbol{\Delta_{\ell_2-TL}}$
NW	0.9613	0.5952	0.3661
NE	0.9484	0.5811	0.3673
SW	0.9475	0.4841	0.4634
SE	0.959	0.5709	0.3881
W	0.9551	0.7718	0.1833
E	0.9452	0.6845	0.2607
SOURCE	0.9588	0.0302	0.9286
Mean (Std)	**0.9528** 0.0065	**0.6146** 0.2390	**0.4225** 0.2410

Table 4.3: Average inner product between $\mathbf{m}_{\text{true}}$ and samples from the 6D-posterior

to the true moment tensor, we also plot some histograms of their relative Euclidean distance (figure 4-11). As it stands out from the plots, the ℓ_2-based posteriors exhibit samples that are much further than to the truth than the TL-based ones. Additionally, the variance appears to be higher in the ℓ_2 case than the TL, confirming a trend already observed in CRPS scores.

We conclude this section by showing some stereonet plots. These plots are meant to represent the fault plane orientations associated with a particular moment tensor. In general, there are two orthogonal planes for each moment tensor and moment tensor inversion does not allow to exclude one of them based on the data. The information

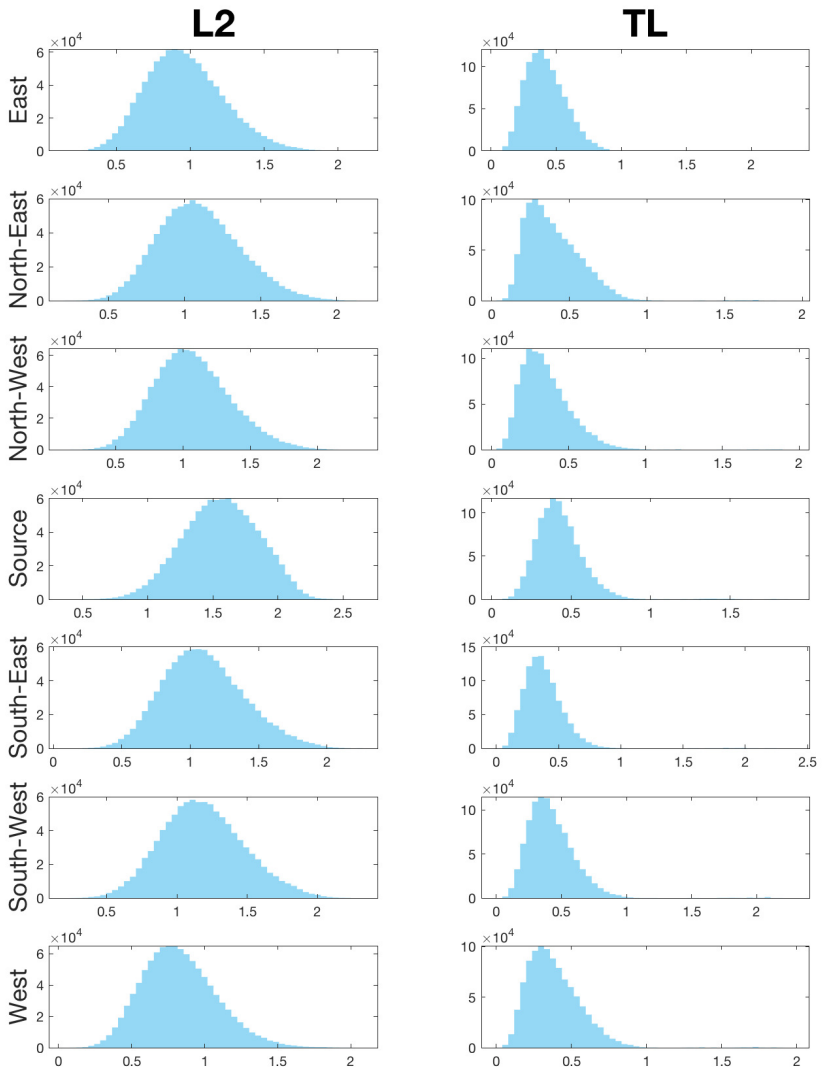


Figure 4-11: Histogram of Euclidean distance between posterior samples and true moment tensor values

about the two planes is visualized as follows: looking down at a hemisphere from above, a dot indicates the location where the normal to one of the two fault planes would intercept the sphere. In Figure 4-12 and we report the results for station NW. The red dots represent the correct answer. It is clear that the TL poles are much more clustered around the correct answer than the ones from the ℓ_2 analysis. The same results can be plotted on a rectangular grid using strike and dip of a pole on the axes

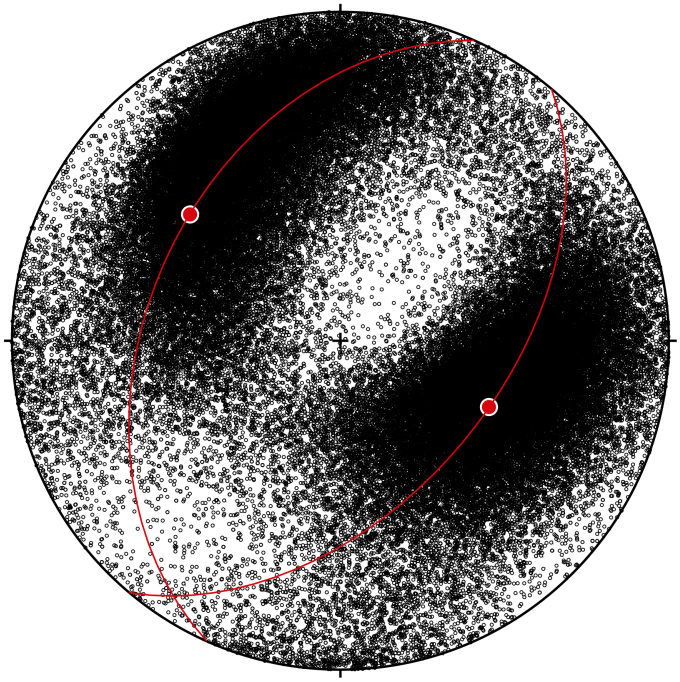


Figure 4-12: Stereonet plot of samples (10^5 samples) from the ℓ_2-based posteriors NW station.

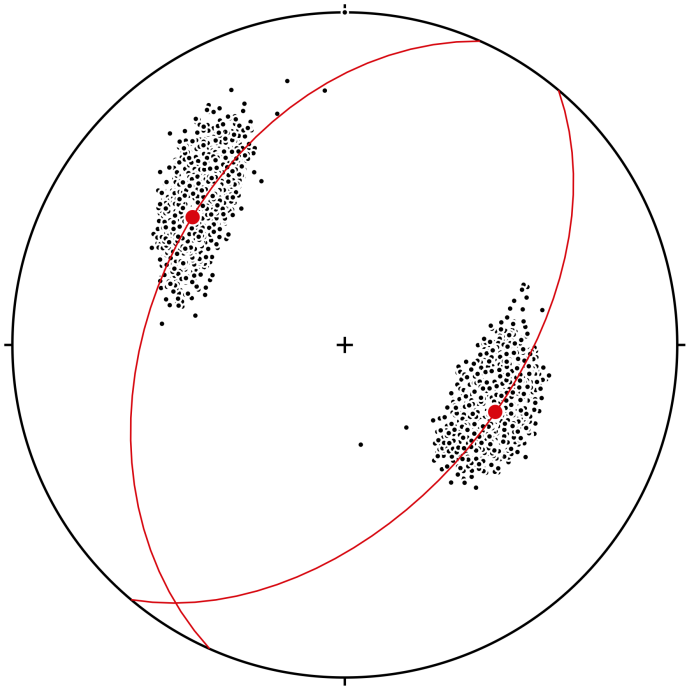


Figure 4-13: Stereonet plot of samples (10^5 samples) from the TL-based posteriors NW station.

(figures 4-14 and 4-15). For completeness we also report the stereonet plots for the

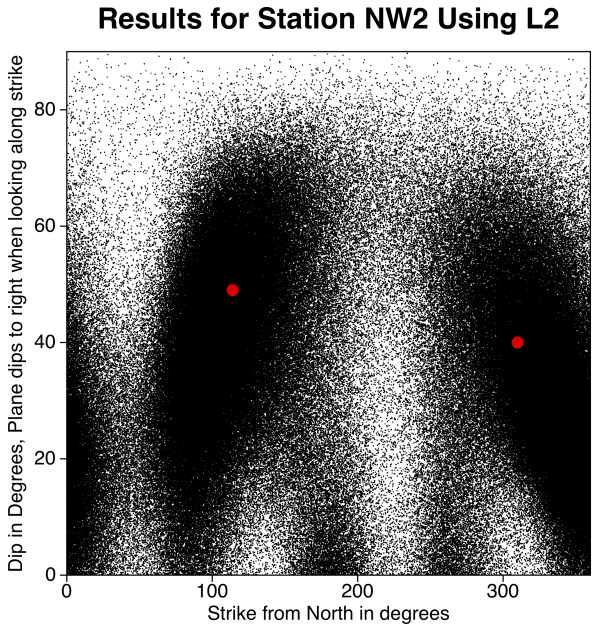


Figure 4-14: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors NW station.

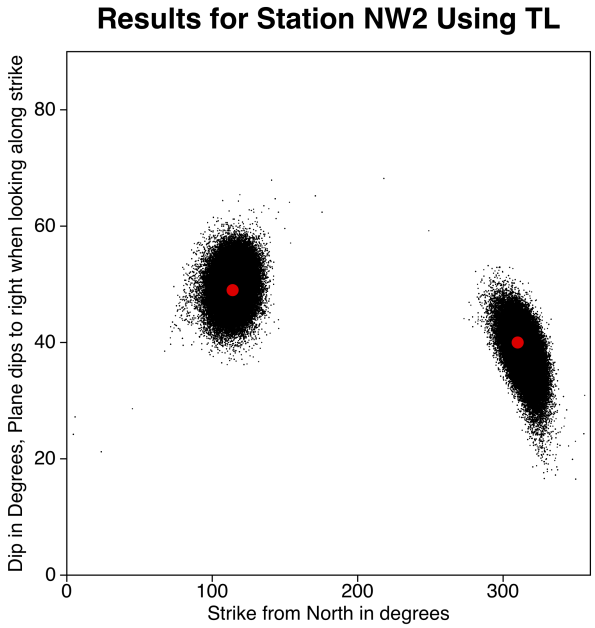


Figure 4-15: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors NW station.

remaining stations 4-16, 4-17, 4-18, 4-19, 4-20, 4-21,4-22, 4-22, 4-24, 4-25, 4-26, 4-27.

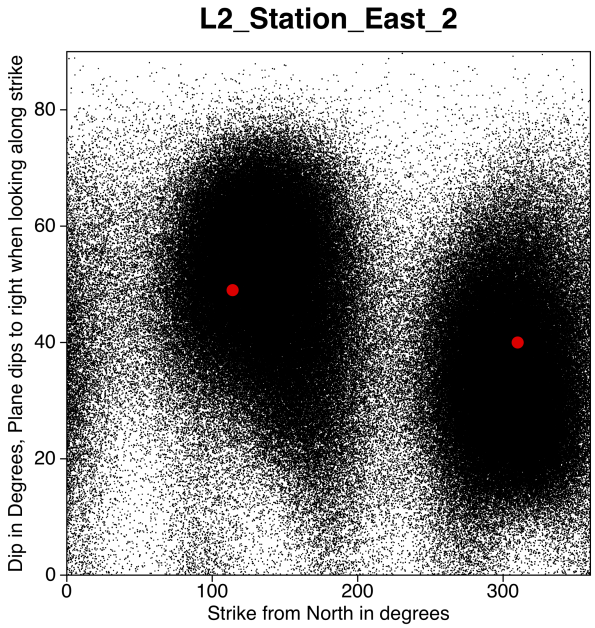


Figure 4-16: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors E station.

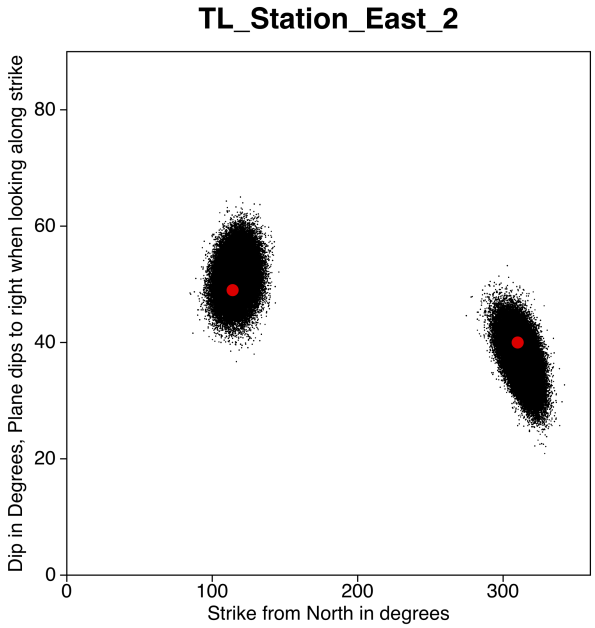


Figure 4-17: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors E station.

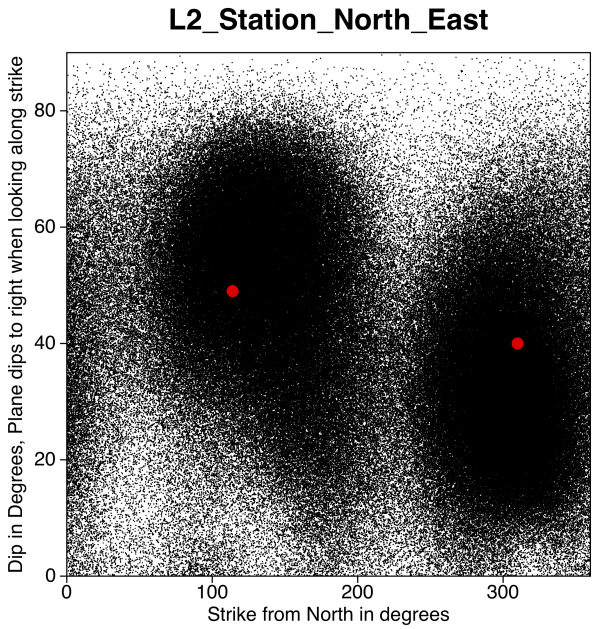


Figure 4-18: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors NE station.

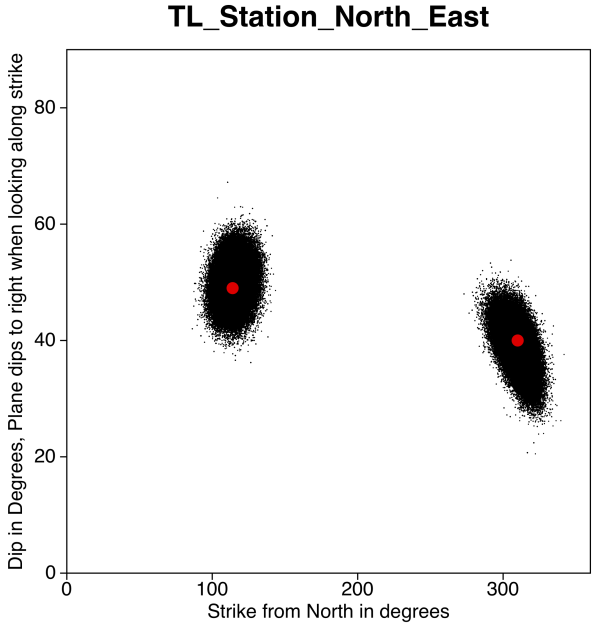


Figure 4-19: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors NE station.

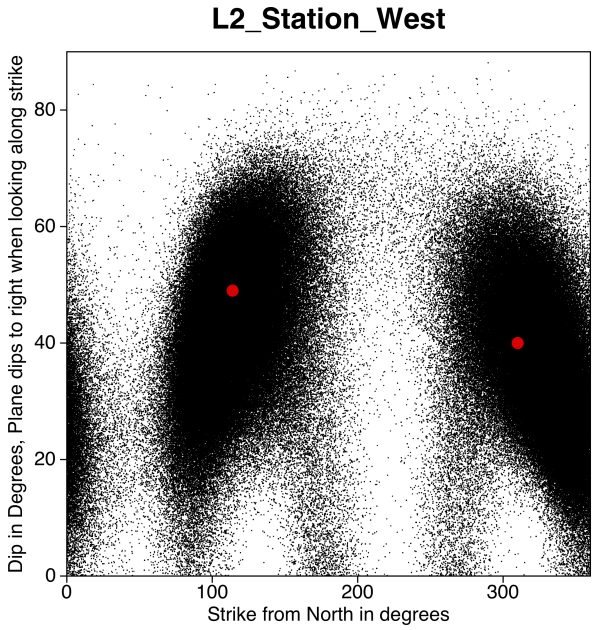


Figure 4-20: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors W station.

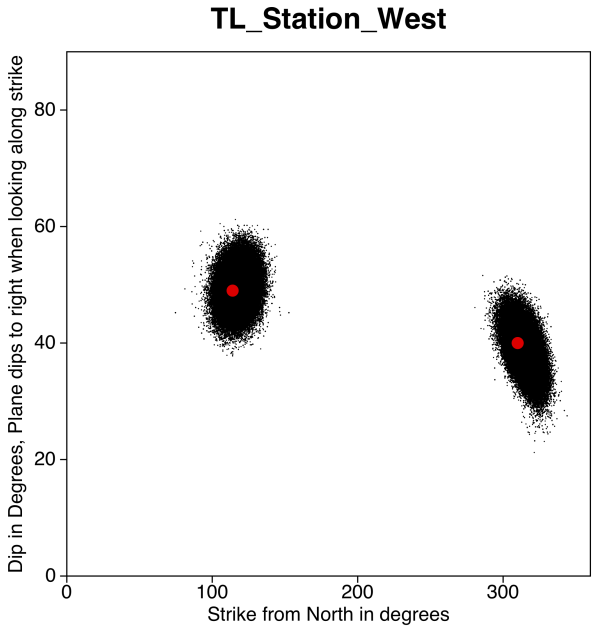


Figure 4-21: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors W station.

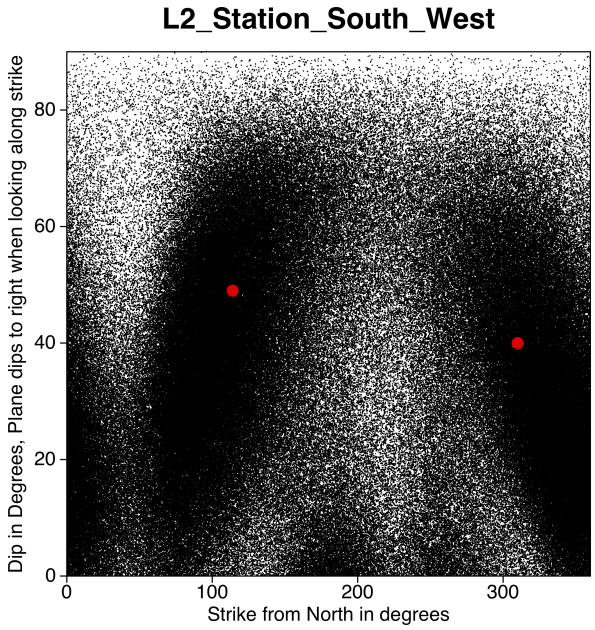


Figure 4-22: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors SW station.

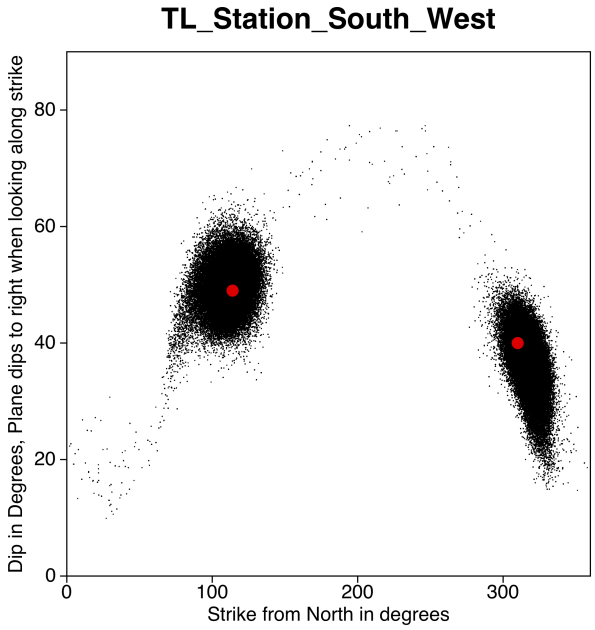


Figure 4-23: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors SW station.

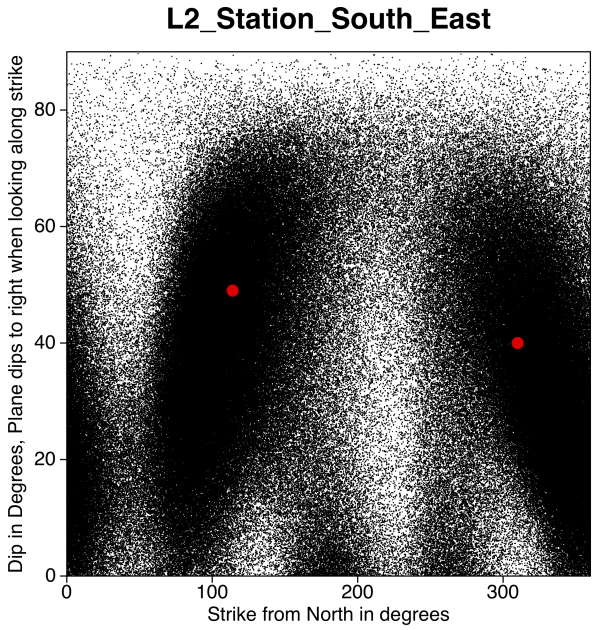


Figure 4-24: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors SE station.

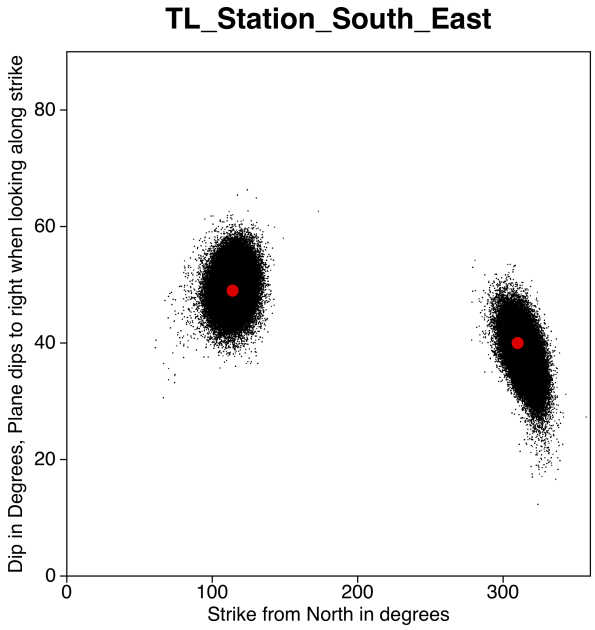


Figure 4-25: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors SE station.

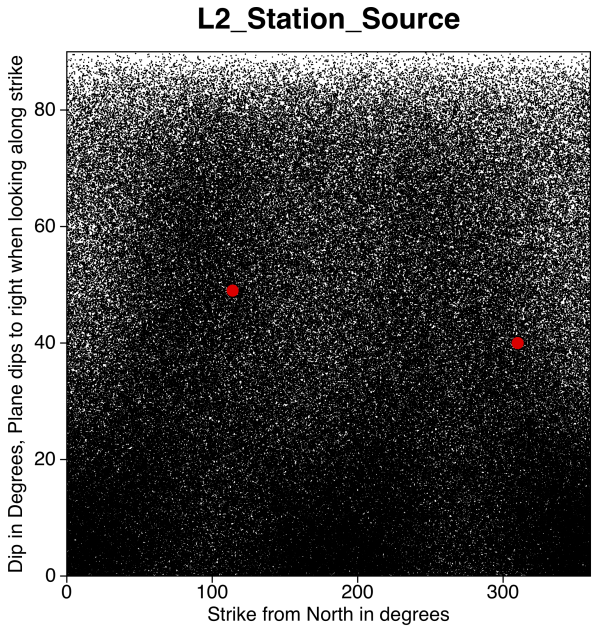


Figure 4-26: Strike-dip plot of samples (10^5 samples) from the ℓ_2-based posteriors Source station.

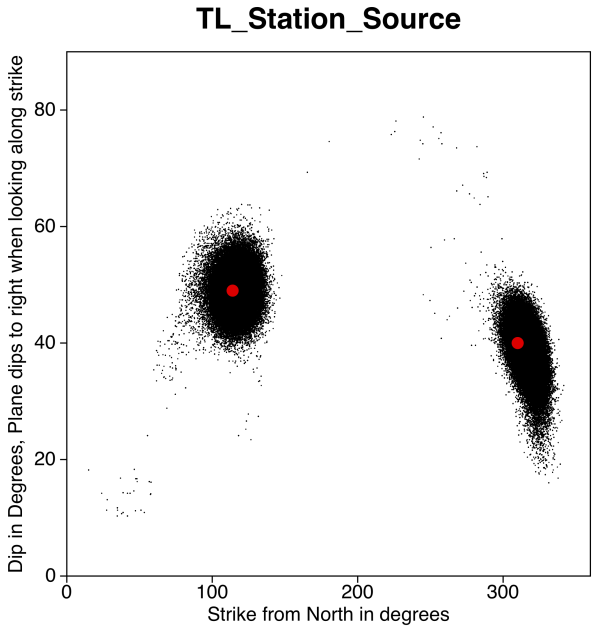


Figure 4-27: Strike-dip plot of samples (10^5 samples) from the TL-based posteriors Source station.

4.3 Impacts of model misspecification on the recovery of double couple vs. non double couple earthquakes

In this section we are concerned with providing an additional point of discussion for a longstanding debate in the seismic community on the modeling and nature of earthquake mechanics. For the most part, earthquakes have long been considered as being generated by shear faulting, i.e., two rigid bodies moving with respect to each other. Mechanically, this motion is traditionally described as the result of the the action of two force-couples (double couple - DC) with zero net torque i.e. no momentum transfer between the source region and the rest of the Earth and volume changes of the source region itself. This description is compatible with the assumptions underlying the classic equations of motion (i.e. $\rho u_{ij} = \sigma_{ij,j}$, with ρ being the regions' density, u_{ij} the displacement in the i,j-th direction and $\sigma_{ij,j}$ is the i,j-th element of the stress tensor acting on the j-th direction). These assumptions also imply that there is no volume change within the source region. More recently, however, it has been shown that real earthquakes could depart from this model. In earthquakes originating from volcanic explosions or other types of non-typical tectonic earthquakes evidence of non-double couple mechanism has emerged [69]. The moment tensor as used throughout this thesis is more general than the purely DC-mechanism and can be in fact decomposed into double couple (DC), isotropic (ISO) and compensated linear vector dipole (CLVD) components, with the letter two being the non-double couple mechanism. The isotropic mechanism originates from a system of forces that act radially (i.e. implosion or explosion) around the source producing a volume change. The CLVD component describes instead mechanisms that have zero-net volume change, but in which two sets of forces act in orthogonal directions: one expanding and the other compressing the source region. A viable interpretation of CLVD is that of a tensile crack with no volume change. In this case, the motion normal to the crack is outward and large. The motion in the orthogonal directions is smaller and inward.

Given a moment tensor, it is possible to decompose it into these three components by means of an eigenvalue decomposition. We refer to [137] for the details of how to perform this calculation.

In this section we analyze the results obtained through our inversion in terms of DC and non-DC percentages. As already mentioned, it is object of debate whether the recovery of any non-DC component in a "natural" or tectonic earthquake is mathematically and physically reliable [35] [143] [2]. Some researchers consider the characterization of isotropic or CLVD components in moment tensor inversion as a byproduct of model misspecification, rather than information actually coming from the data. As a simple test, we analyze whether there is any change in non-DC percentage for events recovered through the TL-based posterior vs. the events recovered through the ℓ_2 distance. We recall that the true moment tensor i.e. the moment tensor used to generate the data is a pure double couple. We report the results in Figure 4-28. The figure clearly shows a reduction in non-DC components of the samples coming from the TL vs. the ℓ_2 distance. Both in terms of CLVD and ISO components of the moment tensor. This results seems to confirm the hypothesis, at least in this case, that the recovery of non-DC mechanisms is mostly linked to the presence of model misspecification rather than coming from the data themselves. As an additional point of view, we also present a pie chart (4-29) where we classified an event as DC if the percentage of DC component was above 60%. From the reported figure it clearly appears that the share of events with a primarily DC component increase drastically among the samples from the TL-based posterior.

4.4 Testing robustness under different focal mechanisms

Up to this point in the chapter, we have tested the performance of the TL distance under inversion with different velocity models. We have however used the same dataset i.e., data generated through the 3D Overthrust model from a single double-couple

Figure 4-28: DC-ISO-CLVD decomposition of samples from the posteriors distributions for each velocity model.

event. In this last section, we instead focus on one layered media model for inversion (NW), while generating data from the Overthrust model using different values of the moment tensor. The objective is to verify that the results described so far are not simply dependent on one specific event. We generate 8 alternative datasets by choosing focal-mechanism that reflect real earthquakes. All of the events have been taken from the Harvard CMT catalogue [66, 38, 39, 26, 40]. We report the information

Figure 4-29: Share of events with a higher than 60% DC component for each velocity model and per TL vs ℓ_2-based posterior.

for the chosen ones in Table 4.4 and 4.5.

These events represent a variety of earthquakes with different percentages of double couple and CLVD components. In Table 4.6 we report the average 1D CRPS scores per each event. It can be seen that for all events except 070886A, the TL represents an improvement in inversion performance. Once again we cannot provide a numerical

Event Name	Mnn	Mne	Mnz	Mee	Mez	Mzz
070886A	-9.933	3	5.247	4.644	-8.325	5.29
070886A - Normalized	-0.739	0.223	0.391	0.346	-0.62	0.394
12487G	-7.28	0.384	-0.945	6.744	1.105	0.536
12487G - Normalized	-0.982	0.052	-0.128	0.91	0.149	0.072
062992L	-1.438	-1.178	0.296	1.413	-0.415	0.025
062992L - Normalized	-0.733	-0.601	0.151	0.72	-0.212	0.013
092904C	-4.75	-1.11	0.002	5.1	0.011	-0.35
092904C - Normalized	-0.909	-0.212	0	0.976	0.002	-0.067
201804051929A	-0.547	-1.25	-0.125	0.523	0.061	0.025
201804051929A - Normalized	-0.398	-0.908	-0.091	0.38	0.044	0.018
201507271812A	0.987	-2	-0.059	-0.676	0.004	-0.311
201507271812A - Normalized	0.425	-0.861	-0.025	-0.291	0.002	-0.134
201511190742A	2.07	-1.11	-0.486	-1.63	-0.076	-0.436
201511190742A - Normalized	0.846	-0.454	-0.199	-0.666	-0.031	-0.178
201511300949A	3.23	-0.651	-0.438	-2.22	-0.325	-1.01
201511300949A - Normalized	0.966	-0.195	-0.131	-0.664	-0.097	-0.302

Table 4.4: List of selected events from the Harvard CMT catalogue [38, 40] - Normalized and unnormalized moment tensor values.

Event Name	Strike	Dip	Rake	Strike	Dip	Rake	DC%	CLVD%
070886A	294	37	156	44	76	55	99.4	0.5
12487G	133	78	178	224	88	12	87.2	12.7
062992L	334	77	173	66	83	13	90.9	9.1
092904C	231	90	0	141	90	180	86.6	13.4
201804051929A	168	84	178	258	88	6	97.4	2.5
201507271812A	191	89	180	281	90	1	73.2	26.8
201511190742A	209	78	179	299	89	12	60.8	39.2
201511300949A	219	75	-173	127	83	-15	43.5	56.5

Table 4.5: List of selected events from the Harvard CMT catalogue [38, 40] - Strike, dip, rake and DC version non-DC percentage.

or intuitive explanation of why this event performed significantly worse than the others. For completeness we report the posterior distributions for each event (Figures 4-30,4-31,4-32,4-33,4-34,4-35,4-36,4-37) together with the associated stereonets with contour plots (Figures 4-38, 4-40, 4-42, 4-44, 4-46, 4-48, 4-50, 4-52, 4-39, 4-41, 4-43, 4-45, 4-47, 4-49, 4-51, 4-53). Generally speaking, except for the first event, the TL contour plots show less biased and less dispersed fault planes recoveries.

Event Name	TL_2	ℓ_2	Δ_{ℓ_2-TL}
070886A	0.7000	0.2027	-0.4973
12487G	0.0282	0.1804	0.1522
062992L	0.0546	0.1702	0.1156
092904C	0.0268	0.1842	0.1574
201804051929A	0.0657	0.1497	0.0840
201507271812A	0.0757	0.1754	0.0997
201511190742A	0.0540	0.1990	0.1450
201511300949A	0.0608	0.1735	0.1127
Mean (Std)	**0.1332** 0.2148	**0.1794** 0.0157	**0.0462** 0.2068

Table 4.6: Average CRPS scores for 1D marginal posteriors

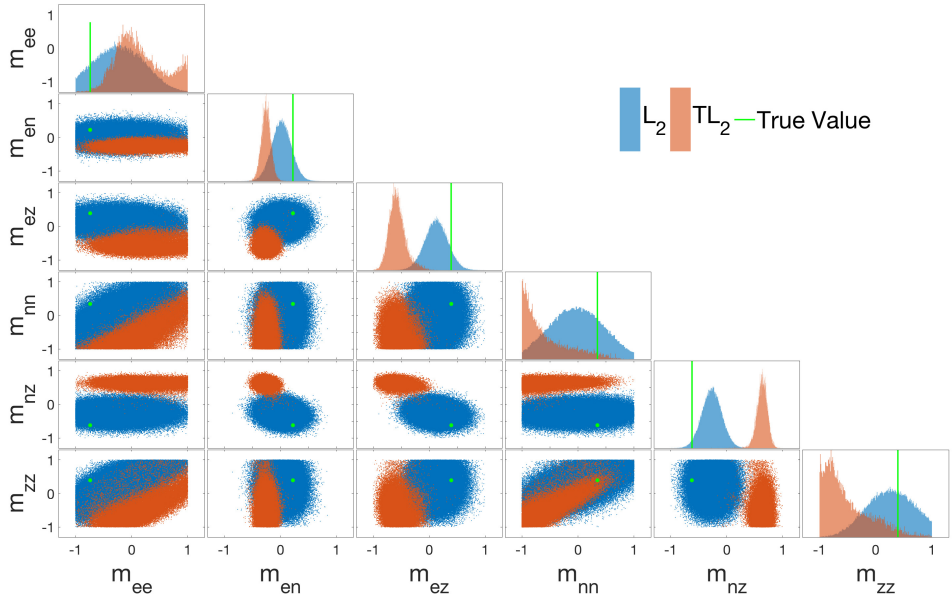


Figure 4-30: Event 070886A: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

4.5 Conclusions

In this chapter we tested the proposed robust Bayesian framework presented in Chapter 3 on a more complex and realistic scenario of moment tensor inversion. The data has been generated using the SEG-EAGE Overthrust velocity model, which is a 3D model. The models used for inversion were instead conceived to be 2D layered-medium approximations of the 3D model. We demonstrated the reliability of the

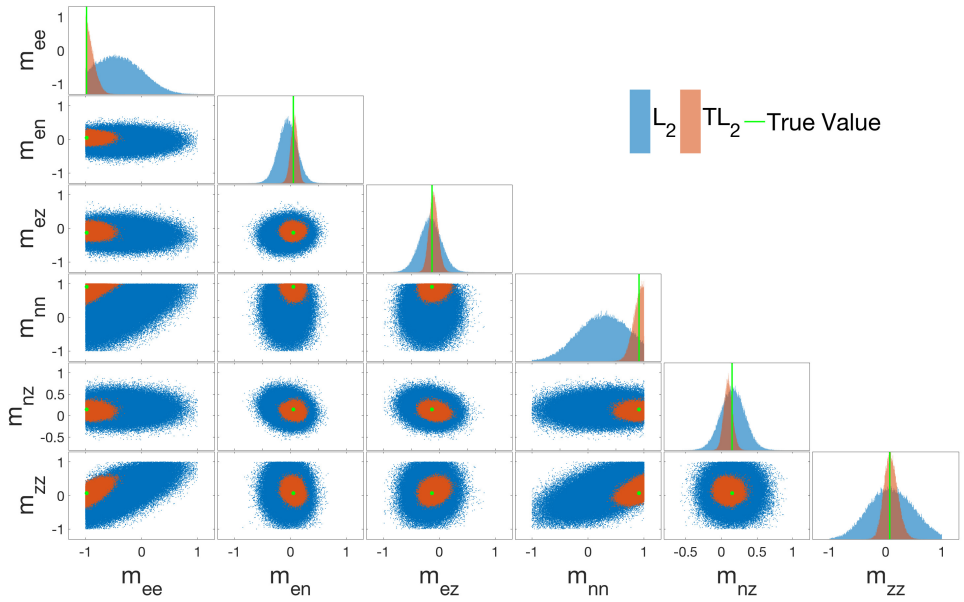


Figure 4-31: Event 12487G: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

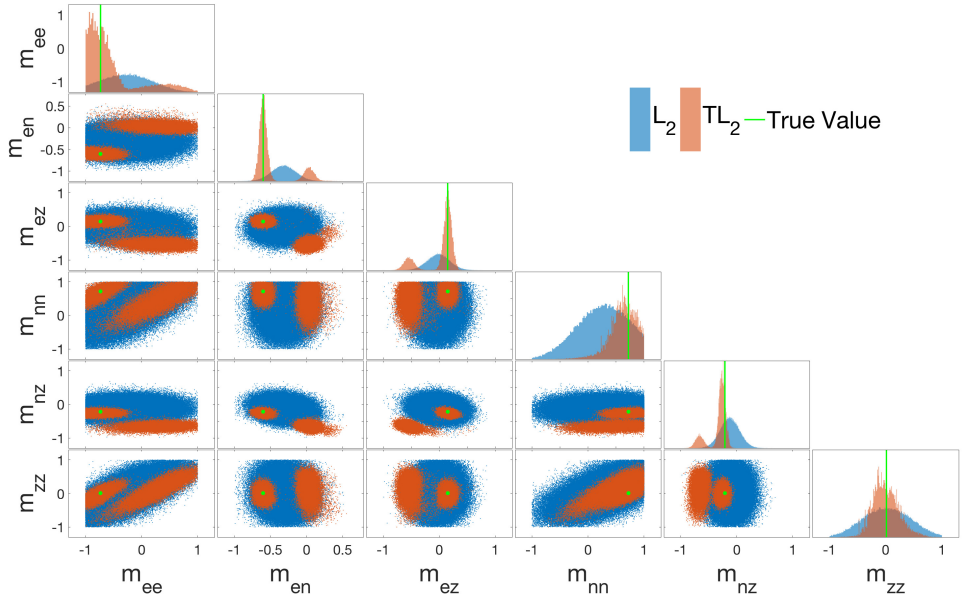


Figure 4-32: Event 062992L: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

methodology in recovering the correct moment tensors under various scenarios of model misspecification as well as source mechanisms. We quantitatively assessed the

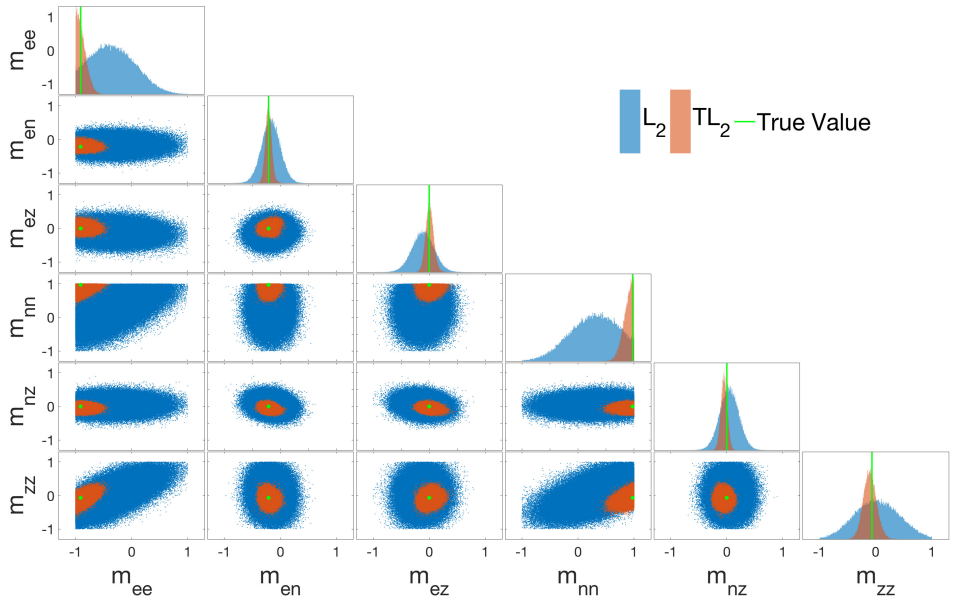


Figure 4-33: Event 092904C: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

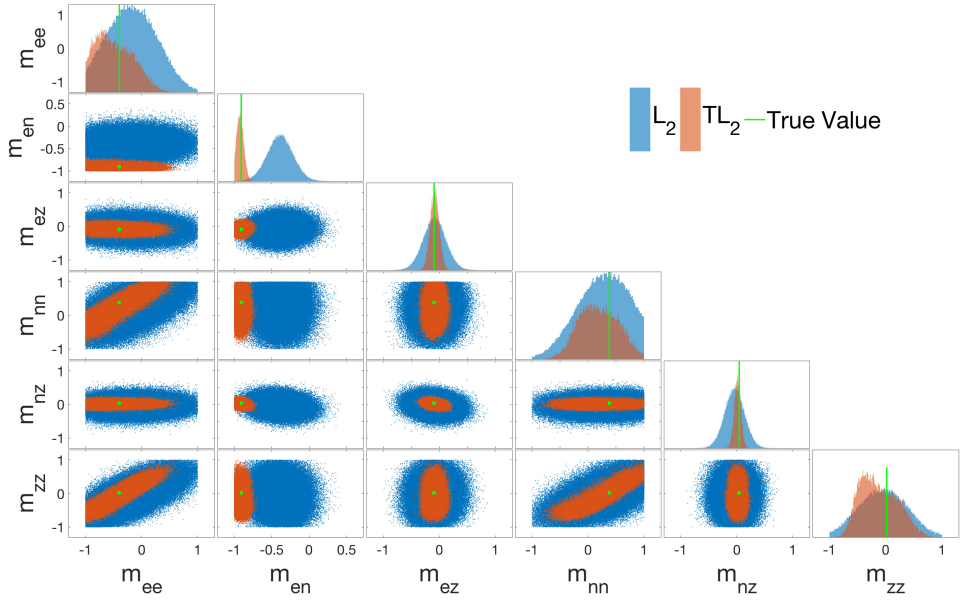


Figure 4-34: Event 201804051929A: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

validity of these results through a number of statistical a geophysical criteria. Finally, we showed how the reduction of the impact of model misspecification on the inversion

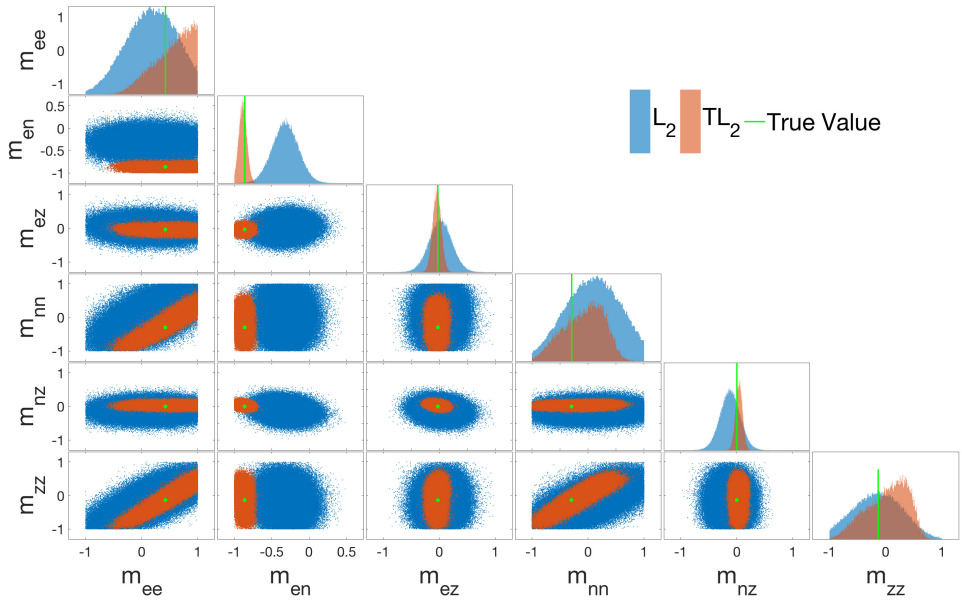


Figure 4-35: Event 201507271812A: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

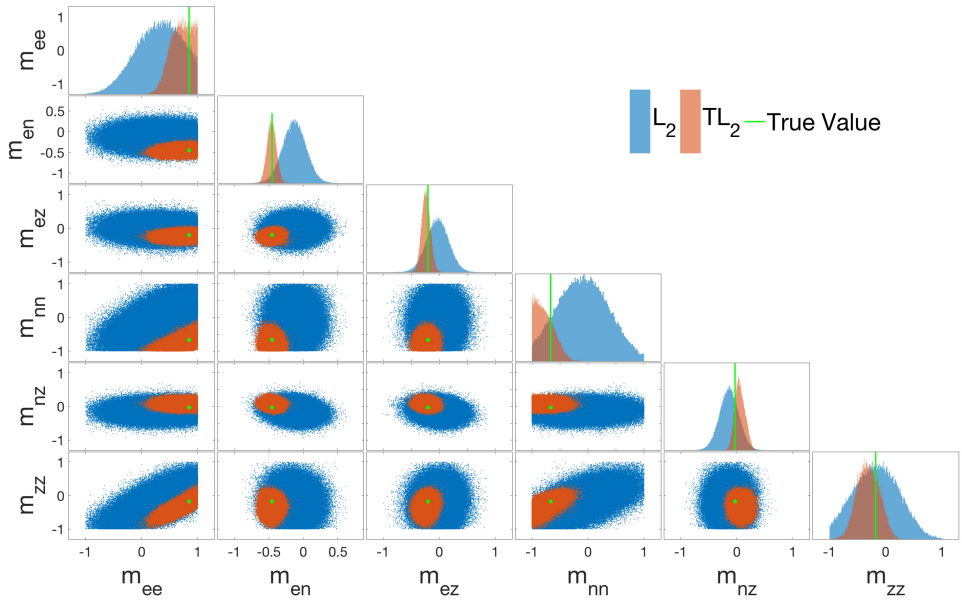


Figure 4-36: Event 201511190742A: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

results has led to a significant decrease in the non double-couple component of the recovered focal mechanisms.

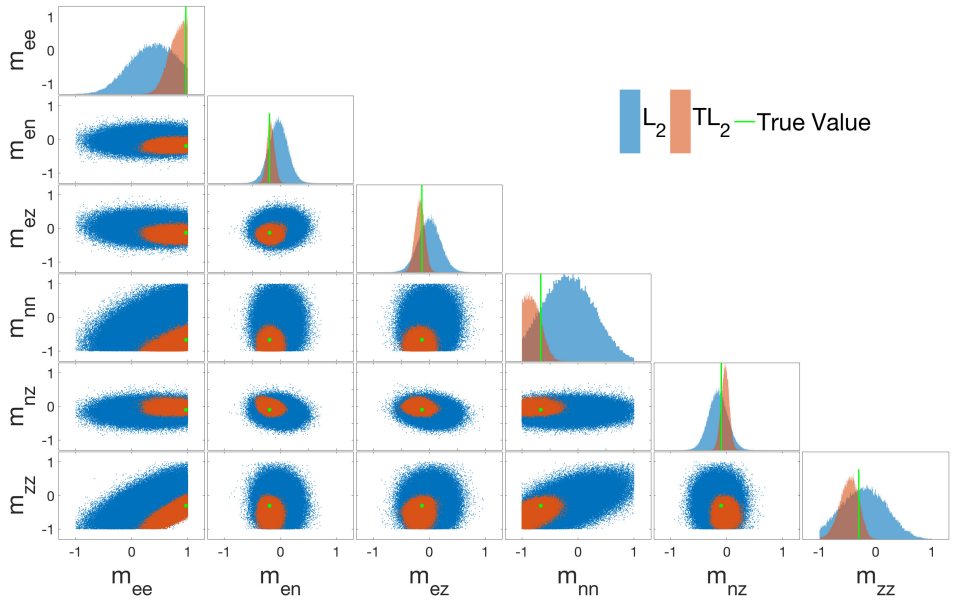


Figure 4-37: Event 201511300949A: matrix plot of the 1D and 2D marginal posteriors for each moment tensor component and misfit measure

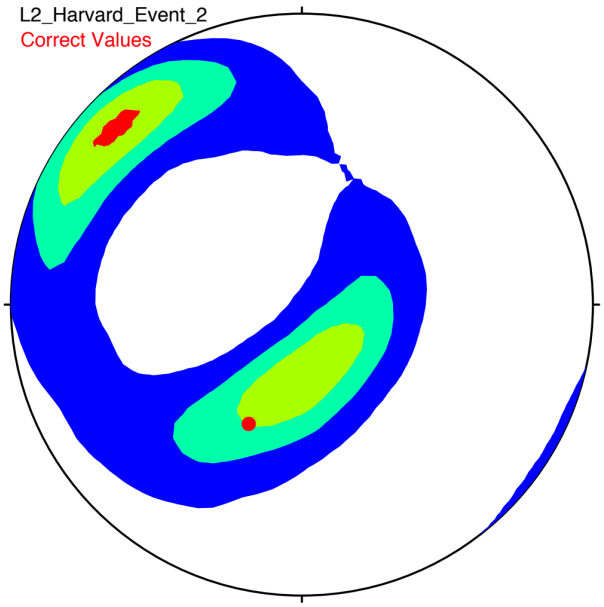


Figure 4-38: Event 070886A: ℓ_2 stereonet plot with contour lines.

Acknowledgements The work presented in this chapter was performed in collaboration with Umair Bin Waheed (KFUPM), SanLinn Kaka (KFUPM), Ben Dia

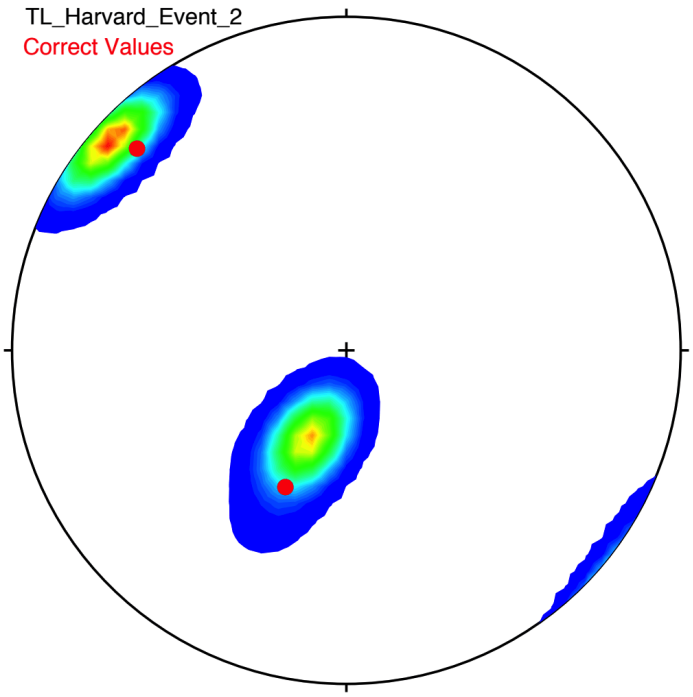


Figure 4-39: Event 070886A: TL stereonet plot with contour lines.

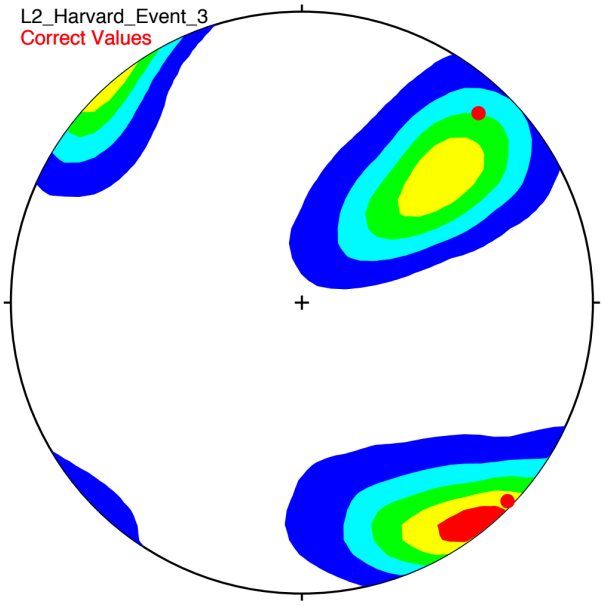


Figure 4-40: Event 12487G: ℓ_2 stereonet plot with contour lines.

(KFUPM) and Chen Gu (MIT).

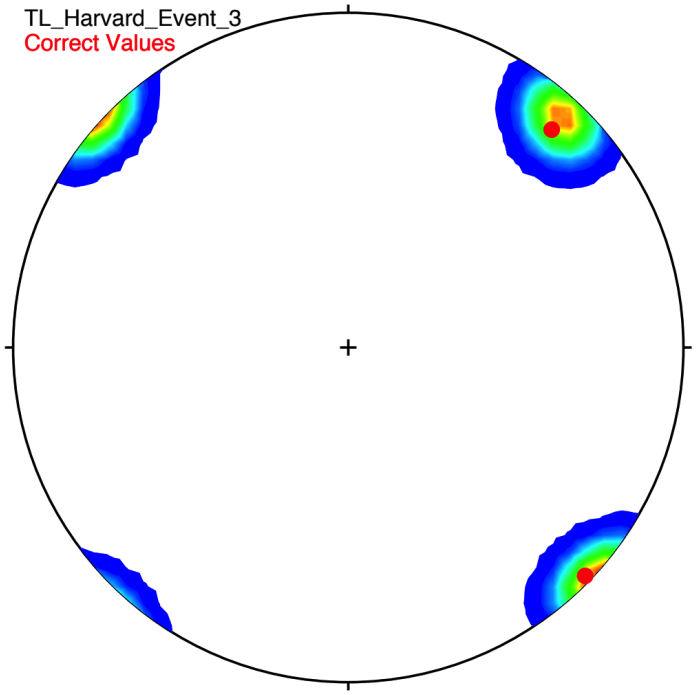


Figure 4-41: Event 12487G: TL stereonet plot with contour lines.

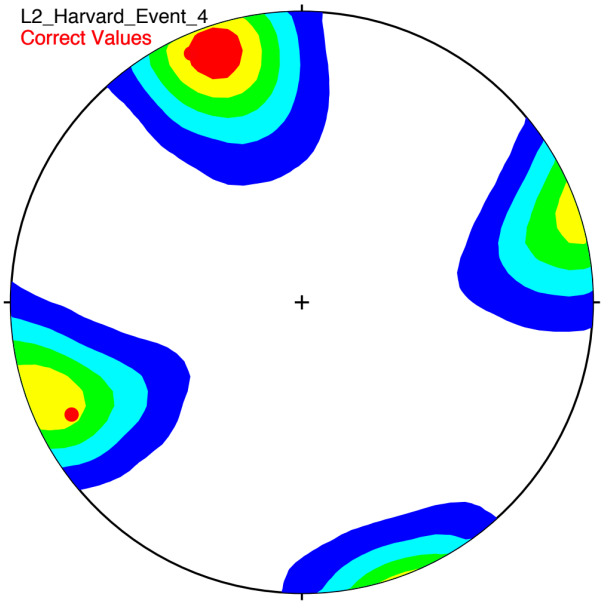


Figure 4-42: Event 062992L: ℓ_2 stereonet plot with contour lines.

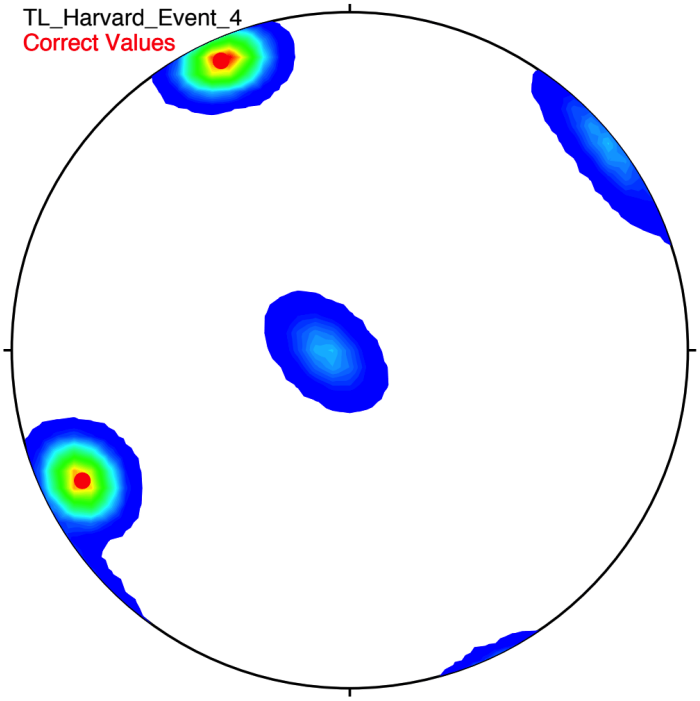


Figure 4-43: Event 062992L: TL stereonet plot with contour lines.

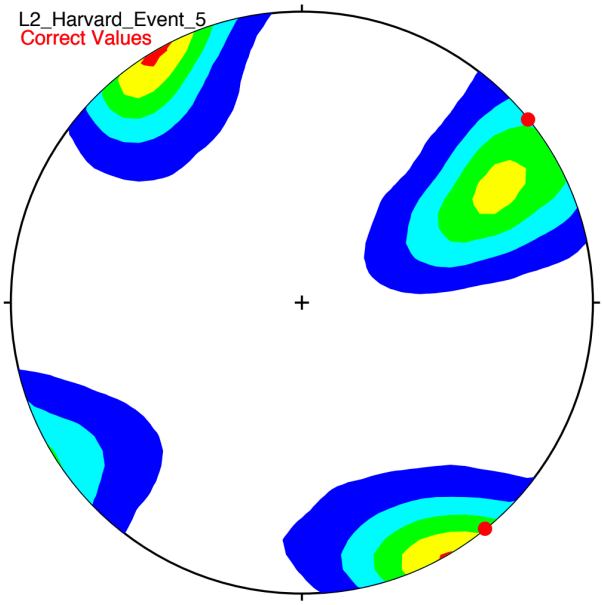


Figure 4-44: Event 092904C: ℓ_2 stereonet plot with contour lines.

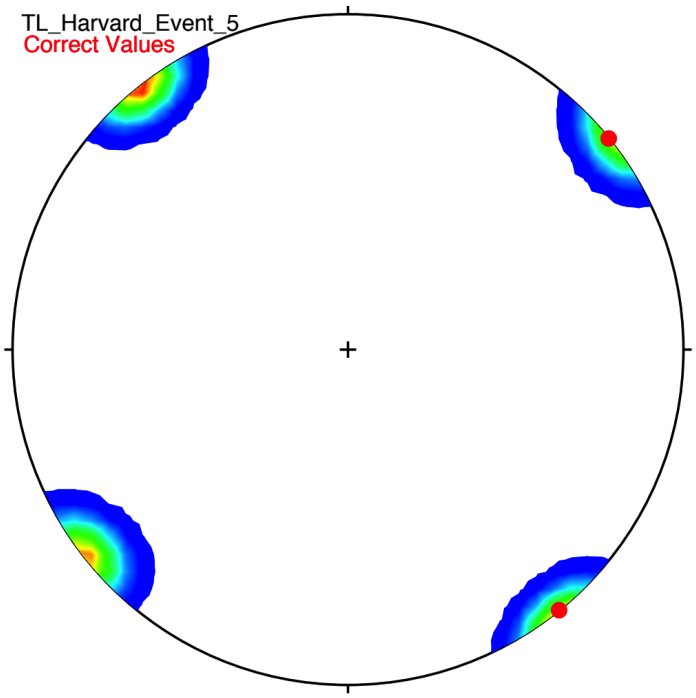


Figure 4-45: Event 092904C: TL stereonet plot with contour lines.

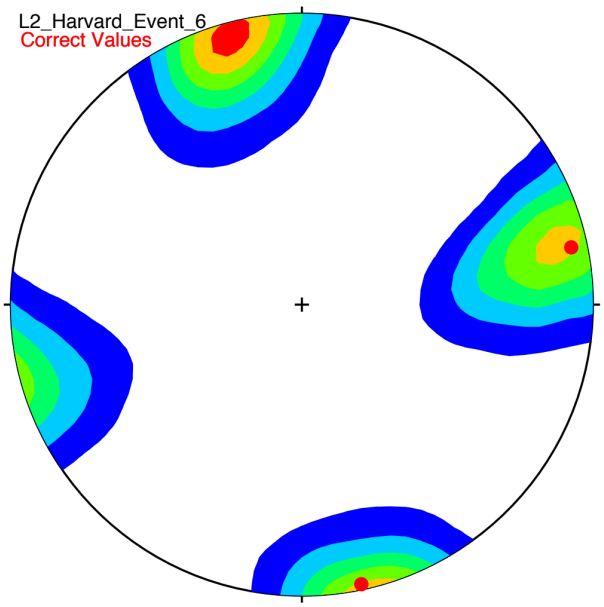


Figure 4-46: Event 201804051929A: ℓ_2 stereonet plot with contour lines.

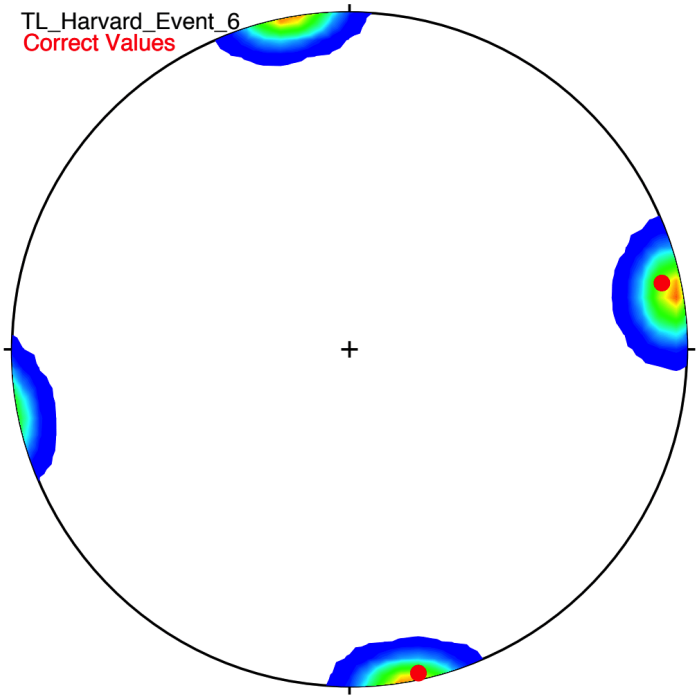


Figure 4-47: Event 201804051929A: TL stereonet plot with contour lines.

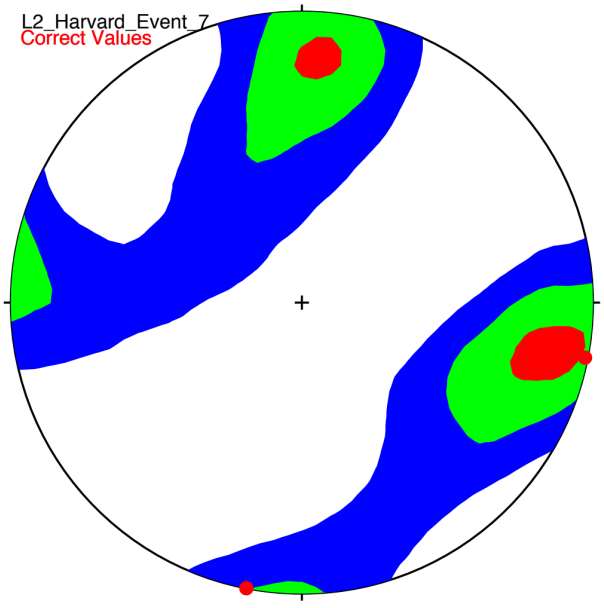


Figure 4-48: Event 201507271812A: ℓ_2 stereonet plot with contour lines.

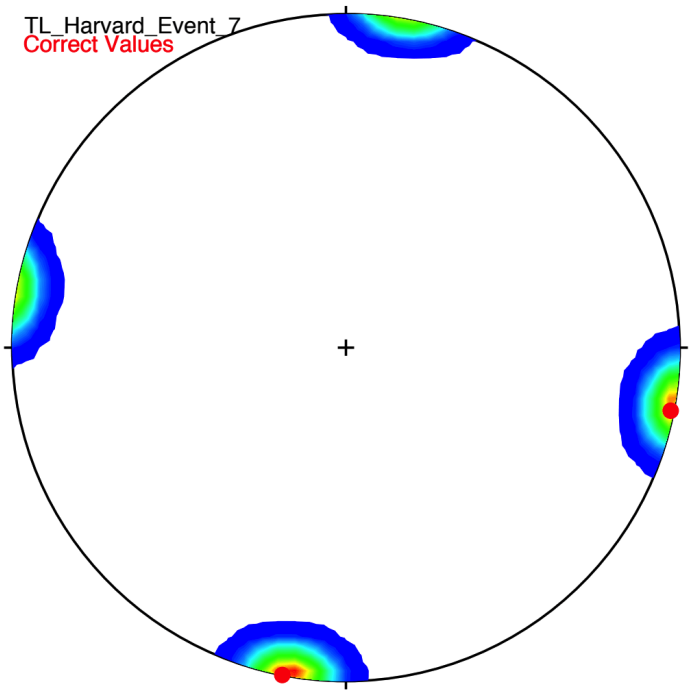


Figure 4-49: Event 201507271812A: TL stereonet plot with contour lines.

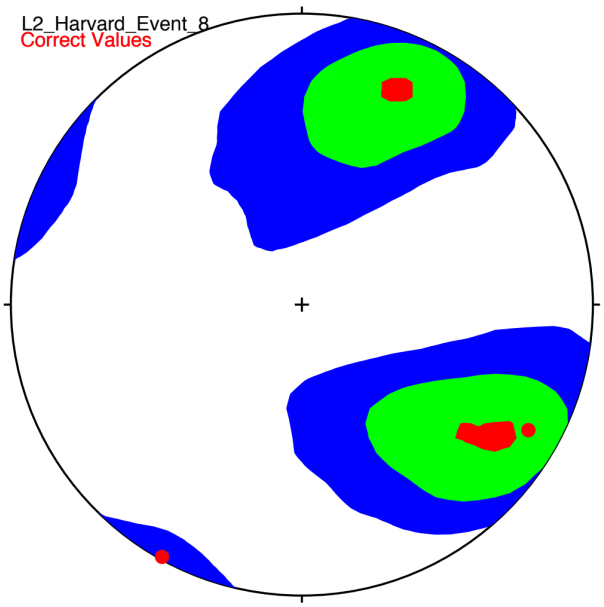


Figure 4-50: Event 201511190742A: ℓ_2 stereonet plot with contour lines.

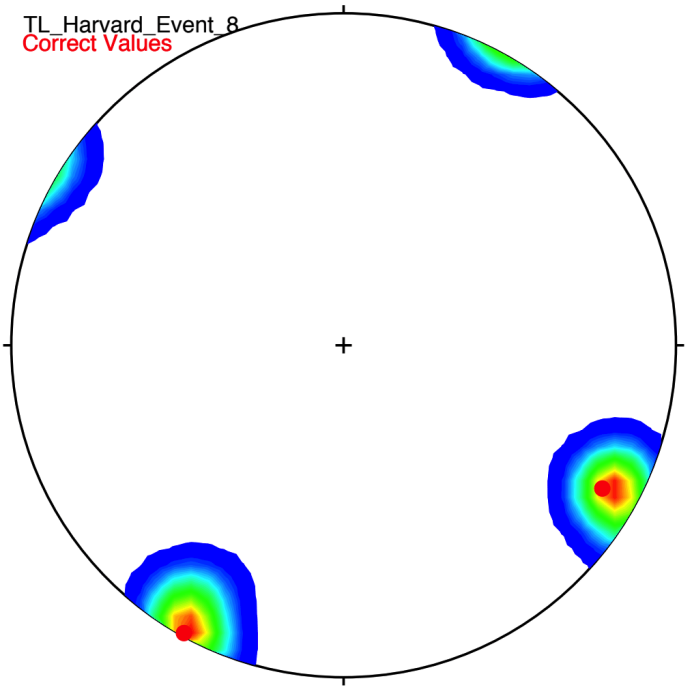


Figure 4-51: Event 201511190742A: TL stereonet plot with contour lines.

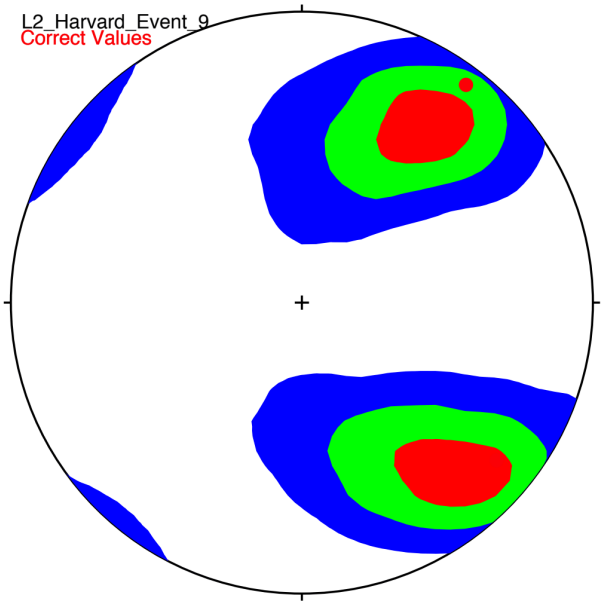


Figure 4-52: Event 201511300949A: ℓ_2 stereonet plot with contour lines.

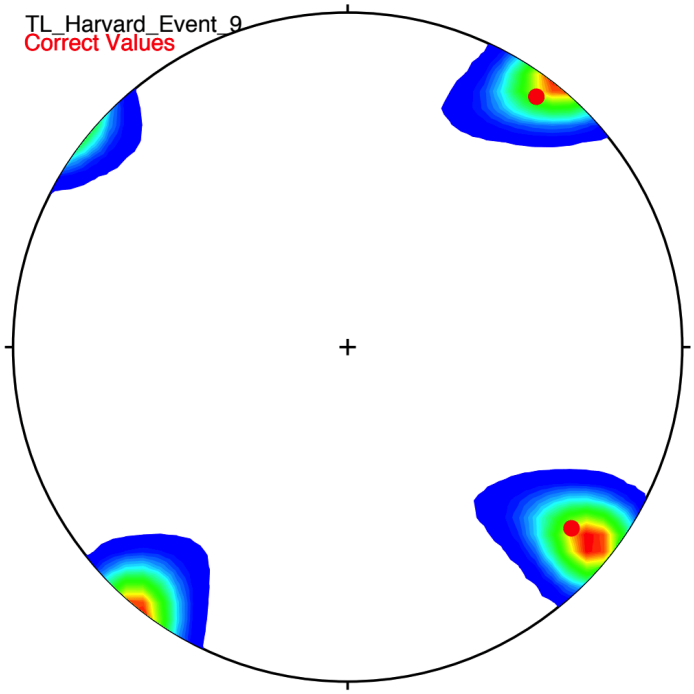


Figure 4-53: Event 201511300949A: TL stereonet plot with contour lines.

Chapter 5

Beyond Gibbs posteriors: statistical properties of the TL distance with additive Gaussian noise

In this chapter we provide some answers to the question of how to characterize the likelihood function associated to a statistical model involving the transport-Lagrangian misfit measure. After defining the problem and several underlying assumptions, we derive an large n closed-form expression for this likelihood. We then provide some geometrical intuition for the problem and describe how this could be used to achieve a tractable approximation of the analytical expression. We close the chapter with some numerical experiments around the statistical behavior of the TL distance.

5.1 A closed form expression for a TL-based likelihood function

5.1.1 The Transport Lagrangian problem setup

In a continuous setting, given two functions, $f, g : [T_0, T] \to \mathbb{R}^n$, the transport-Lagrangian problem is formulated as follows [128, 129]:

$$\mathrm{TL}_p^\lambda = \min_{\pi \in \Pi} \int_{T_0}^{T} \int_{T_0}^{T} \quad (\lambda |x - y|^p + |f(x) - g(y)|^p) \pi(x, y) dx dy$$

where:

$$\Pi = \left\{ \pi : [T_0, T]^2 \to \mathbb{R}_{\geq 0} \text{ s.t. } \forall x, y \in [T_0, T] : \int_{T_0}^{T} \pi(x, z) dz = \int_{T_0}^{T} \pi(z, y) dz = \frac{1}{T - T_0} \right\}$$

The set Π contains all the densities 2D $\pi(x, y)$ such that their marginals are uniform distributions over the interval $[T_0, T]$. In a discrete setting (i.e., functions f, g discretized at n points x_i, y_j) the couplings π can be restricted to being all possible permutations $\boldsymbol{\sigma}$ of n elements $\mathrm{Perm}(n)$ (i.e., vertices of the Birkhoff polytope) [101]:

$$\mathrm{TL}_p^\lambda = \min_{\boldsymbol{\sigma} \in \mathrm{Perm}(n)} \sum_{i=1}^{n} |x_i - x_{\sigma(i)}|^p + \lambda |f(x_i) - g(x_{\sigma(i)})|^p = \tag{5.1}$$

$$= \min_{\boldsymbol{\sigma} \in \mathrm{Perm}(n)} ||\mathbf{x} - \mathbf{x}_\sigma||_p^p + ||f(\mathbf{x}) - g(\mathbf{x}_\sigma)||_p^p \tag{5.2}$$

There are $n!$ possible permutations for a set of n elements, which implies the summation above (5.2) can only take on $n!$ values. We call each of these values:

$$Z_k = ||\mathbf{x} - \mathbf{x}_{\boldsymbol{\sigma}_\mathbf{k}}||_p^p + \lambda ||f(\mathbf{x}) - g(\mathbf{x}_{\boldsymbol{\sigma}_\mathbf{k}})||_p^p, \tag{5.3}$$

where k stands for the specific assignment induced by the permutation $\boldsymbol{\sigma}_k$. The TL

problem therefore reduces to:

$$\mathrm{TL}_p^{\lambda} = \min\{Z_k\}_{k=1}^{n!} \tag{5.4}$$

5.1.2 Introducing randomness and signal-model interpretation

We now interpret the $g(y), f(x)$ as signals $g(t_j), f(t_i)$ and rewrite the indices x_i, y_j as time indices t_i, t_j. We then consider $g(t_i)$ as the observation collected at time t_i, while $f(t_i)$ is the model prediction from a deterministic model with *added Gaussian noise*, i.e., $f(t_i) = \mu_{f(t_i)} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \rho)$. With these premises, we can rewrite the Z_k in more compact notation:

$$Z_k = a_k + \lambda \sum_{i=1}^{n} |g(t_{\sigma_k(i)}) - f(t_i)|^p \tag{5.5}$$

where:

$$a_k = ||\mathbf{t} - \mathbf{t}_{\boldsymbol{\sigma}_\mathbf{k}}||_p^p$$

We note that the $f(t_i)$ are independent Gaussian random variables, each with a (potentially) different mean, but same variance. Without loss of generality, we assume $\lambda = 1$. We also choose $p = 2$, which restricts the scope of our discussion to what is a standard choice for cost functions based on Euclidean metrics. With this choice of p we can state that:

$$X_l = \left(\frac{g(t_j) - f(t_i)}{\rho}\right)^2 \sim \chi^2_{1,\gamma=\mu_l^2} \text{ with } i, j: \ 1, \ldots, n; \ l: 1, \ldots, n^2 \text{ and } l = i + (j-1) * n \tag{5.6}$$

Each X_l is therefore a non-central[1] chi-squared random variable with 1 degree of freedom and $\gamma = \mu_l^2$, where $\mu_l = (\mu_{f(t_i)} - g(t_j))/\rho$. Although there are $n!$ possible values that Z_k can take, there are only n^2 possible values that $|g(t_j) - f(t_i)|^2$ can assume. We can therefore express all the Z_k as a linear system in X_l:

$$Z = A + \rho^2 BX \tag{5.7}$$

where:

$$\begin{aligned} Z &= [Z_1, \ldots, Z_{n!}] \in \mathbb{R}_{\geq 0}^{n!} \\ A &= [a_1, \ldots, a_{n!}] \in \mathbb{R}_{\geq 0}^{n!} \\ X &= [X_1, \ldots, X_{n^2}] \in \mathbb{R}_{\geq 0}^{n^2} \\ B &\in \{\{0,1\}^{n! \times n^2} \text{s.t. } B\mathbb{1}^{n^2} = n\mathbb{1}^{n!}\} \end{aligned}$$

$\mathbb{1}$ is vector of 1, B is a "selector" matrix which simply selects and sums the values of X_l ($||g(t_j) - f(t_i)||^2$) that contribute to the Z_k associated to permutation k. Because the B matrix only has n entries equal to 1 per row, each $\frac{Z_k - a_k}{\rho^2}$ is then a non-central chi-squared random variable with n degrees of freedom and γ specified as follows:

$$\frac{Z_k - a_k}{\rho^2} \sim \chi^2_{n,\gamma} \quad \text{with:} \quad \gamma = \sum_{l=1}^{n^2} b_{k,l}\mu_l^2 \tag{5.8}$$

Where b_k is the k-th row of B.

[1] To completely specify a non-central chi-squared random variable two parameters are needed: one indicating the degrees of freedom and another one indicating its non-centrality, which we refer to as γ. This is equal to the sum of the squared means of each of the normal random variables being squared ad summed.

5.1.3 Finding the minimum of a set of dependent and non-identically-distributed random variables

While a large body of literature exists on order statistics, these are normally derived for a set of i.i.d random variables. Nonetheless, according to [30] and [44] the CDF of the minimum $\min(Z)$ of a collection of random variables $Z \equiv (Z_1, \ldots, Z_{n!})$ (with no independence or distributional assumptions) can be expressed by calculating the probability that *at least* one of the Z_k is less or equal to z. In particular:

$$\mathbb{P}(\min(Z) \leq z) = \sum_{k=1}^{n!} (-1)^{k-1} S_k \tag{5.9}$$

$$S_k = \sum_{\substack{\tau \\ \tau \in \mathcal{S}^{\{1,2,\ldots n!\}} \\ |\tau| = k}} \mathbb{P}(Z_{\tau_1} \leq z, Z_{\tau_2} \leq z, \ldots, Z_{\tau_k} \leq z) \tag{5.10}$$

where $\mathcal{S}^{\{1,2,\ldots n!\}}$ is the power set of the $n!$ elements. The term S_k has to be interpreted as the sum of the joint CDFs of all subsets of cardinality k of the $\{Z_1, \ldots Z_{n!}\}$ random variables Conceptually the formula just expresses the need to avoid "double counting" when dealing with events that "overlap" (dependency) and is in fact a version of the inclusion-exclusion principle [70]. For example, in the two dimensional case we would have:

$$\begin{aligned}
\mathbb{P}(\min(Z) \leq z) &= S_1 - S_2 \\
S_1 &= \mathbb{P}(Z_1 \leq z) + \mathbb{P}(Z_2 \leq z) \\
S_2 &= \mathbb{P}(Z_1 \leq z, Z_2 \leq z)
\end{aligned}$$

The probability of the minimum of two random variables being less or equal to z is calculated as the sum of the probabilities of each of the two random variables being less than z individually, minus (to avoid double counting) the probability of both of them being less than z simultaneously.

The problem therefore reduces to calculating the joint CDF of each subset of Z_k-s

needed to compute the above sum. In general this task can be rather complex, but in our case we can obtain the joint CDFs of any subset of $\{Z_k\}_1^{n!}$ by exploiting their linear dependence to the X_l and their particular covariance structure.

5.1.4 Obtaining the joint CDF of all possible subsets of $\{Z_k\}_{k=1}^{n!}$

We have already stated that each X_l is a non-central chi-squared random variable with one degree of freedom. We have also already concluded that given the linearity of the relationship (5.7), the $\frac{Z_k - a_k}{\rho^2}$ are non-central chi-squared random variables with n degrees of freedom and parameter γ as specified in (5.8). While this information is useful in characterizing the nature of the Z_k-s as random variables, a closed form expression for their multivariate CDF is not available.

In order to proceed further, it is convenient to characterize their large n behavior. Given the nature of the application we are dealing with, we can safely assume that $n > 100$ (discretization points). Additionally, even though the $\{\chi^2_{1,\lambda=\mu_l^2}\}_1^{n^2}$ are not all independent from each other, each row of $B \cdot X$ is a sum of n independent $\chi^2_{1,\lambda=\mu_l^2}$. Recalling the definition:

$$X_l = \left(\frac{g(t_j) - f(t_i)}{\rho} \right)^2, \tag{5.11}$$

the particular nature of the assignment problem makes it such that a Z_k will never be a sum of X_l that originate from the same $f(t_i)$. Since the $f(t_i)$-s are independent by construction, then we will always be dealing with a sum of independent random variables.

This consideration and the large values of n allow us to apply the Lyapunov central limit theorem (see applicable conditions in appendix A) and approximate the distribution of the $B \cdot X$ with a normal distribution. Subsequently, it becomes trivial to derive the distribution of any subset of Z_k (linear combination of X_l, see (5.7)).

For the entire set of $\{Z_k\}_{k=1}^{n!}$, we would have:

$$\boldsymbol{\mu} = A + \rho^2 \cdot B \cdot \mathbb{E}(X) \tag{5.12}$$

$$\boldsymbol{\Sigma} = \rho^4 B \cdot \mathbb{C}\text{ov}(X) \cdot B^T \tag{5.13}$$

Let us now consider a generic subset of cardinality k of the power set $\mathcal{S}^{\{1,2,\ldots,n!\}}$ containing the random variable $\{Z_{\tau_1}, Z_{\tau_2}, \ldots, Z_{\tau_k}\}$. Each of these subsets will be distributed according to a multivariate Gaussian with mean and covariance matrix defined as:

$$\boldsymbol{\mu}_{\boldsymbol{\tau_1,\tau_2,\ldots,\tau_k}} = \begin{bmatrix} a_{\tau_1} \\ a_{\tau_2} \\ \cdots \\ a_{\tau_k} \end{bmatrix} + \rho^2 \cdot \begin{bmatrix} — & b_{\tau_1} & — \\ — & b_{\tau_2} & — \\ & \cdots & \\ — & b_{\tau_k} & — \end{bmatrix} \cdot \mathbb{E}(X) \tag{5.14}$$

$$\boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k} = \rho^4 \begin{bmatrix} — & b_{\tau_1} & — \\ — & b_{\tau_2} & — \\ & \cdots & \\ — & b_{\tau_k} & — \end{bmatrix} \cdot \mathbb{C}\text{ov}(X) \cdot \begin{bmatrix} — & b_{\tau_1} & — \\ — & b_{\tau_2} & — \\ & \cdots & \\ — & b_{\tau_k} & — \end{bmatrix}^T \tag{5.15}$$

We can now complete the description by making the mean and covariance matrices for the vector X explicit. As discussed before, the X_l are non-central chi squared random variables. Therefore, referring to definitions in (5.6), we have:

$$\mathbb{E}(X) = \mathbb{1} + [\mu_1^2, \ldots, \mu_{n^2}^2]^T$$

For the covariance matrix, based on the discussion made before on the independence-dependence structure of the X_l (5.6) we have:

$$\mathbb{C}\text{ov}(X_{l'}, X_{l''}) = 0 \iff i_{l'} \neq i_{l''} \tag{5.16}$$

$$\mathbb{C}\text{ov}(X_{l'}, X_{l''}) \neq 0 \iff i_{l'} = i_{l''} \tag{5.17}$$

In particular the non-zero entries of the matrix will be either on the diagonal ($i_{l'} = i_{l''}$ and $j_{l'} = j_{l''}$, which means $l' = l''$) or those X_l that share the same $f(t_i)$, but whose mean has been offset by a different constant $g(t_j)$. While the value on the diagonal entries is simply the variance of each X_l, we report the value of the latter case in the appendix B only. The important aspect to remember is that, given the nature of the assignment problem, the off diagonal non-zero entries of $\mathbb{C}\text{ov}(X)$ will be nullified by the zero entries of B in the product $B \cdot \mathbb{C}\text{ov}(X) \cdot B^T$, and we can thus simplify the discussion by considering $\mathbb{C}\text{ov}(X)$ as a diagonal matrix:

$$\mathbb{C}\text{ov}(X) \triangleq \mathbb{I}_{n^2 \times n^2}\big(2\mathbb{1} + 4[\mu_1^2, \ldots, \mu_{n^2}^2]^T\big)$$

5.1.5 From the CDF to the PDF of Z

Based the above discussion we are able to propose an approximate analytic expression for the CDF of $\min(Z)$ when n is sufficiently large. Given:

$$\text{TL}_2 = \min\{Z_k\}_1^{n!},$$

by calling $\mathbf{\Phi}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(z)$ the CDF of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we have that:

$$\begin{aligned}
\mathbb{P}(\min(Z) \leq z) &= \sum_{k=1}^{n!} (-1)^{k-1} S_k \\
S_k &= \sum_{\substack{\tau \in \mathcal{S}^{\{1,2,\ldots n!\}} \\ |\tau| = k}} \Phi_{\boldsymbol{\mu}_{\tau_1,\tau_2,\ldots,\tau_k}, \boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k}}(z)
\end{aligned} \tag{5.18}$$

Before proceeding with the derivation of an expression for the PDF of Z, we note that a closed form expression for a multivariate normal CDF does not exist. At best, we can express $\Phi_{\boldsymbol{\mu}_{\tau_1,\tau_2,\ldots,\tau_k}, \boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k}}(z)$ as:

$$\Phi_{\boldsymbol{\mu}_{\tau_1,\tau_2,\ldots,\tau_k}, \boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k}}(z) = \iint \cdots \int_{-\infty}^{z} \phi(z_{\tau_1}, z_{\tau_2}, \ldots, z_{\tau_k})\, dz_{\tau_1} dz_{\tau_2} \ldots dz_{\tau_k}, \tag{5.19}$$

where ϕ is the joint density function of a multivariate Gaussian with mean and covariance $\boldsymbol{\mu}_{\tau_1,\tau_2,\ldots,\tau_k}, \boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k}$. It is well know that in order to get the PDF $(p(\cdot))$ of a random variable it is sufficient to differentiate its CDF. In our case this requires taking the total derivative with respect to z:

$$p(z) = \sum_{k=1}^{n!} (-1)^{k-1} S_k \tag{5.20}$$

$$S_k = \sum_{\substack{\tau \\ \tau \in \mathcal{S}^{\{1,2,\ldots n!\}} \\ |\tau|=k}} \frac{d}{dz} \iint \cdots \int_{-\infty}^{z} \phi(z_{\tau_1}, z_{\tau_2}, \ldots, z_{\tau_k})\, dz_{\tau_1} dz_{\tau_2} \ldots dz_{\tau_k} \tag{5.21}$$

The derivative of the integral in (5.21) can be calculated via the multivariate chain rule. In particular, we want to compute, for the function $\phi : \mathbb{R}^k \to \mathbb{R}$, the derivative of the scalar function $H : \mathbb{R} \to \mathbb{R}$ given by the iterated integral

$$H(z) = \int_{-\infty}^{z} \cdots \int_{-\infty}^{z} \phi(\tau_1, \ldots, \tau_k) d\tau_1 \ldots d\tau_k.$$

To do this, we note that H can be expressed as the composition $H = H_2 \circ H_1$, where:

$$H_1 : \mathbb{R} \to \mathbb{R}^k, \quad z \mapsto [z, \ldots, z]^T \in \mathbb{R}^k,$$

and

$$H_2 : \mathbb{R}^k \to \mathbb{R}, \quad (s_1, ..., s_k) \mapsto \int_{-\infty}^{s_1} \cdots \int_{-\infty}^{s_k} \phi(\tau_1, \ldots, \tau_k) d\tau_1 \ldots d\tau_k.$$

We then see that at each point z and $i = 1, ..., k$, we have $(\partial H_{1,i})'(z) = 1$, and moreover that for any $s \in \mathbb{R}^k$ and $i = 1, ..., k$,

$$\frac{\partial H_2}{\partial s_i}(s) = \int_{-\infty}^{s_1} \cdots \int_{-\infty}^{s_k} \phi(\tau_1, \ldots, \tau_{i-1}, s_i, \tau_{i+1}, ..., \tau_k) d\tau_1 \ldots d\tau_{i-1} d\tau_{i+1} \ldots d\tau_k,$$

where the notation in the above display means integration over all but the i-th variable (i.e., $k-1$ variables). Using the chain rule, we notice that for any $z \in \mathbb{R}$, denoting the

vector $\mathbf{z} := [z, \ldots, z]^T$ the derivative $H'(z)$ is given by:

$$\begin{aligned} H'(z) &= \sum_{i=1}^{k} \frac{\partial H_2}{\partial z_i}(\mathbf{z}) = \\ &= \sum_{i=1}^{k} \Big\{ \int_{-\infty}^{z} \cdots \int_{-\infty}^{z} \phi(\tau_1, \ldots, \tau_{i-1}, z, \tau_{i+1}, ..., \tau_k) d\tau_1 \ldots d\tau_{i-1} d\tau_{i+1} \ldots d\tau_k \Big\} = \\ &= \sum_{i=1}^{k} \phi_{\tau_i} \Phi_{\tau_{-i}|\tau_i} \end{aligned}$$

By substituting this expression back into (5.21), we obtain:

$$p_{\min(Z)}(z) = \sum_{k=1}^{n!} (-1)^{k-1} T_k \tag{5.22}$$

$$T_k = \sum_{\substack{\tau \in \mathcal{S}^{\{1,2,\ldots n!\}} \\ |\tau| = k}}^{\tau} \sum_{i=1}^{k} \phi_{\tau_i} \Phi_{\tau_{-i}|\tau_i} \tag{5.23}$$

In order to facilitate the interpretation of the above formula and compare it to the case where Z_i are independent, we focus on the $n = 2$ case, which allows an effective pictorial representation of the following expression:

$$\begin{aligned} p_{\min\{Z_1, Z_2\}}(z) = p_{Z_1}(z) + p_{Z_2}(z) + \\ - p_{Z_1}(z)\mathbb{P}(Z_2 \leq z | Z_1 = z) - p_{Z_2}(z)\mathbb{P}(Z_1 \leq z | Z_2 = z) \end{aligned}$$

This expression does not rely on the specific nature of the distribution of the single Z_i, but expresses a general principle represented in Figure (5-1). For a set of dependent and non-identically distributed random variables, we have that the probability of the minimum being equal to a specific value z is given by the sum of the probability of each random variable being equal to z minus (or plus) additional terms that take into account the fact that when one of the random variables of the set is fixed to z the others may not be less than z since this would imply that z is not the actual minimum. This concept is mathematically expressed, in the 2D case, through the product: $p_{Z_1}(z)\mathbb{P}(Z_2 \leq z | Z_1 = z)$ and $p_{Z_2}(z)\mathbb{P}(Z_1 \leq z | Z_2 = z)$. In the special case

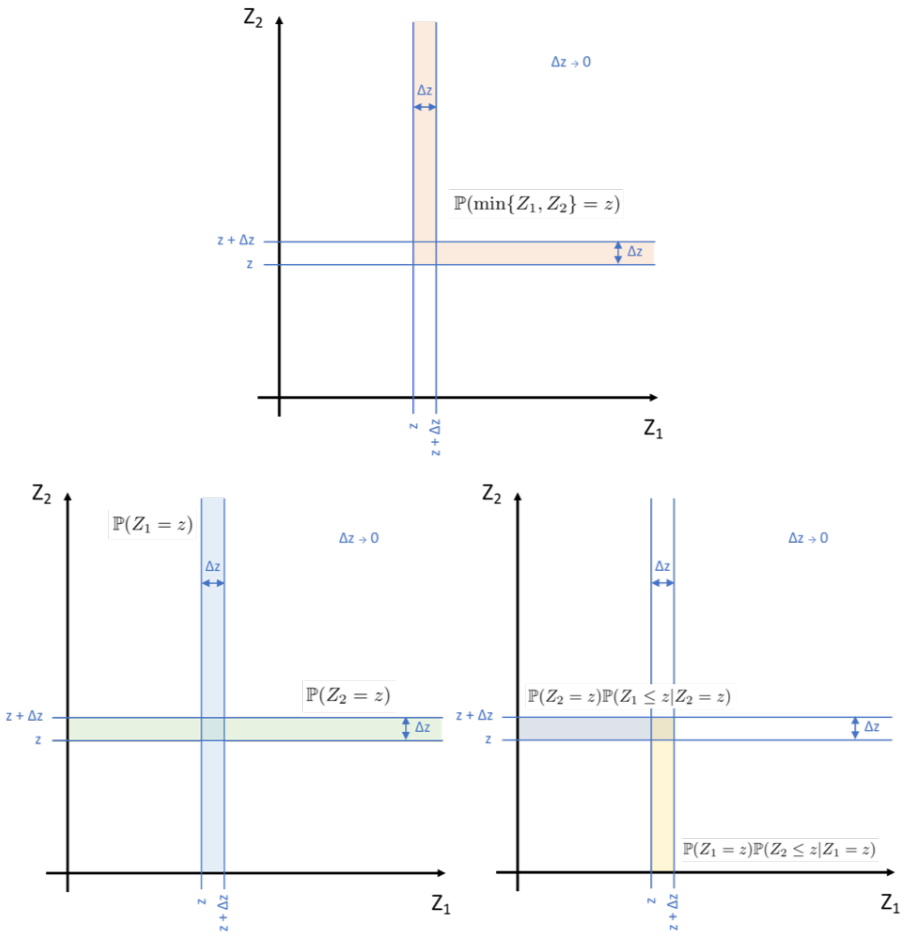


Figure 5-1: Construction of $P(\min\{Z_1, Z_2\} = z)$

when the Z_k are independent, then, by definition, $\mathbb{P}(Z_i \leq z | Z_j = z) = \mathbb{P}(Z_i \leq z)$ and in the 2-D case:

$$p_{\min\{Z_1,Z_2\}}(z) = p_{Z_1}(z) + p_{Z_2}(z) - p_{Z_1}(z)\mathbb{P}(Z_2 \leq z) - p_{Z_2}(z)\mathbb{P}(Z_1 \leq z)$$

This is consistent with the general formula for order statistics of independent random variables [30]:

$$\begin{aligned} p_{\min\{Z_1,Z_2\}}(z) &= \frac{d}{dz}[1 - (1 - \mathbb{P}(Z_1 \leq z))(1 - \mathbb{P}(Z_2 \leq z))] = \\ &= (1 - \mathbb{P}(Z_1 \leq z))p_{Z_2}(z) + (1 - \mathbb{P}(Z_2 \leq z))p_{Z_1}(z) = \\ &= p_{Z_1}(z) + p_{Z_2}(z) - p_{Z_1}(z)\mathbb{P}(Z_2 \leq z) - p_{Z_2}(z)\mathbb{P}(Z_1 \leq z). \end{aligned}$$

5.1.6 Conclusion

Based on the above discussion we can therefore express the PDF of:

$$\mathrm{TL}_2 = \min\{Z_k\}_{k=1}^{n!},$$

as:

$$p_{\mathrm{TL}_2}(z) = \sum_{k=1}^{n!}(-1)^{k-1} \sum_{\substack{\tau \in \mathcal{S}^{\{1,2,\dots n!\}} \\ |\tau|=k}} \sum_{i=1}^{k} \phi_{\tau_i} \Phi_{\tau_{-i}|\tau_i} \tag{5.24}$$

This expression is generalizable to any set of random variable (not necessarily normally distributed) as long as their CDFs and PDFs are available (numerically or analytically).

While the approximation is almost exact for large n, the main impediment to any practical use of this expression is the number of terms implicit in the double summations over $n!$ and $2^k \ \forall k : 1 \dots n!$. In the next sections we therefore investigate the numerical behavior of the newly derived TL-likelihood and propose a number of pathways to a tractable approximation.

5.2 A geometric viewpoint

One of the crucial impediments to obtaining a tractable expression for the TL-based likelihood is the fact that the Z_k Gaussian random variables are not independent. In particular, while we can easily write:

$$\mathbb{P}(\mathrm{TL}_2 \leq z) = 1 - \mathbb{P}(Z_1 \geq z, Z_2 \geq z, \dots, Z_{n!} \geq z),$$

it then is impossible to factor the second term into a product of unidimensional CDFs. Even if that were possible, there would still be a tractability issue arising from the extremely high number of factors ($n!$). A first step towards a simplification of the problem can be that of exploiting its low-dimensional statistical structure. In fact,

although we characterized the random nature of the Z_k-s alone, these variables can be also expressed via deterministic coupling with a set of n i.i.d. mean-zero Gaussian random variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \rho^2\mathbb{I})$ (see equations (5.6) and (5.7)). We can therefore write:

$$\mathbb{P}(\mathrm{TL}_2 \leq z) = 1 - \mathbb{P}(Z_1 \geq z, Z_2 \geq z, \ldots, Z_{n!} \geq z) = \tag{5.25}$$

$$= 1 - \int_z^\infty \cdots \int_z^\infty p(Z = z|\epsilon_1, \ldots, \epsilon_n) p(\epsilon_1, \ldots, \epsilon_n) d\epsilon_1 \ldots d\epsilon_n \tag{5.26}$$

$$= 1 - \int_z^\infty \cdots \int_z^\infty p(Z = z|\epsilon_1, \ldots, \epsilon_n) \prod_{i=1}^n p(\epsilon_i) d\epsilon_1 \ldots d\epsilon_n \tag{5.27}$$

The density $p(Z = z|\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ is in fact degenerate and simply represents the aforementioned deterministic coupling between Z and X:

$$Z = A + \rho^2 BX$$

which in turn implies a set of quadratic relationships between Z and $\boldsymbol{\epsilon}$ since X is a vector containing non-central chi-squared distributions, and therefore $X = X(\boldsymbol{\epsilon})$. We express these relationships through the notation $Z = \mathcal{P}(\boldsymbol{\epsilon})$ (which can be through as a system of quadratic equations in $\boldsymbol{\epsilon}$) and reformulate the above integral as:

$$\mathbb{P}(\mathrm{TL}_2 \leq z) = 1 - \int\limits_{z \leq \mathcal{P}(\boldsymbol{\epsilon})} \prod_{i=1}^n p(\epsilon_i) d\epsilon_1 d\epsilon_2 \ldots d\epsilon_n$$

At this point the main challenge shifts towards determining the nature of the region described by the relation $z \leq \mathcal{P}(\boldsymbol{\epsilon})$ in the n-dimensional $\boldsymbol{\epsilon}$ space. A variable Z_k can be expressed as:

$$Z_k = a_k + \sum_i^n (g(t_{\sigma(i)}) - \mu_{f(t_i)} - \epsilon_i)^2$$

This equation defines an n-dimensional sphere in the $\{\epsilon_i\}_{i=1}^n$ coordinate space with radius $\sqrt{Z_k - a_k}$. We call $\mathcal{S}_k^{\mathbf{c},z}$ the sphere defined by the equation above and all its interior points and where $\mathbf{c} = [g(t_{\sigma(1)}) - \mu_{f(t_1)}, \ldots, g(t_{\sigma(n)}) - \mu_{f(t_n)}]$ is its center and z the radius. The region $z \leq \mathcal{P}(\boldsymbol{\epsilon})$, for a given z, will be defined by the complement of

the union of the $n!$ $\mathcal{S}_k$ spheres:

$$\mathcal{W}(z) = \{\boldsymbol{\epsilon} \in \mathbb{R}^n : \mathcal{P}(\boldsymbol{\epsilon}) \geq z\} = \mathbb{R}^n \setminus \bigcup_{k=1}^{n!} \mathcal{S}_k^{\mathbf{c},z}$$

This result is again challenging to apply in practice without some sort of approximation in the limit of $n \to \infty$. For this reason we provide some additional facts regarding the geometry just defined. To simplify the notation we call $g(t_i) = g_i$ and $\mu_{f(t_i)} = \mu_i$.

Proposition 5.1. *Given a set of $n!$ hyperspheres with centers and radii defined respectively as:*

$$\mathbf{c}_k = [g_{\sigma(1)} - \mu_1, g_{\sigma(2)} - \mu_2, \ldots, g_{\sigma(n)} - \mu_n] \in \mathbb{R}^n \tag{5.28}$$

$$r_k^2 = z - a_k \tag{5.29}$$

there exists a sphere on which all of the $\{\mathbf{c_k}\}_{k=1}^{n!}$ lie, whose center $\mathbf{C}$ and radius R are, respectively:

$$\mathbf{C} = \left[\frac{1}{n}\sum_{i=1}^{n} g_i - \mu_1, \frac{1}{n}\sum_{i=1}^{n} g_i - \mu_2, \ldots, \frac{1}{n}\sum_{i=1}^{n} g_i - \mu_n\right] \in \mathbb{R}^n \tag{5.30}$$

$$R^2 = \sum_{i}^{n}\left(g_i - \frac{1}{n}\sum_{i=1}^{n} g_i\right)^2 \tag{5.31}$$

Proof. Let us assume a sphere exists on which all of the $\{\mathbf{c_k}\}_{k=1}^{n!}$ lie and that its center $\mathbf{C}$ is:

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} g_i - [\mu_1, \mu_2, \ldots, \mu_n]$$

Then it must hold that the distance between each of the $\{\mathbf{c_k}\}_{k=1}^{n!}$ and the point $\mathbf{C}$ is

constant. In particular:

$$||\mathbf{c}_k - \mathbf{C}||_2^2 = |||[g_{\sigma^k(1)} - \mu_1, g_{\sigma^k(2)} - \mu_2, \ldots, g_{\sigma^k(n)} - \mu_n] - \mathbf{C}||_2^2 = \tag{5.32}$$

$$= |||[g_{\sigma^k(1)} - \mu_1, g_{\sigma^k(2)} - \mu_2, \ldots, g_{\sigma^k(n)} - \mu_n] - \frac{1}{n}\sum_{i=1}^{n} g_i + [\mu_1, \mu_2, \ldots, \mu_n]||_2^2 = \tag{5.33}$$

$$= |||[g_{\sigma^k(1)}, g_{\sigma^k(2)}, \ldots, g_{\sigma^k(n)}] - \sum_{i=1}^{n} g_i||_2^2 = \sum_{i=1}^{n}\left(g_i - \sum_{i=1}^{n} g_i\right)^2 = R^2 \tag{5.34}$$

where the last term is independent of k, i.e., any specific permutation. □

With this result, the region $\mathcal{W}(z)$ can be understood as having a particular structure, which could be exploited, under certain circumstances, to facilitate the calculation of the integral:

$$\int_{\mathcal{W}(z)} \prod_{i=1}^{n} p(\epsilon_i) d\epsilon_1 d\epsilon_2 \ldots d\epsilon_n \tag{5.35}$$

Figures 5-2 and 5-3 represent the configuration of $n! = 6$ spheres ($n = 3$) $\{\mathcal{S}_k^{\mathbf{c},z}\}_{k=1}^{6}$ for the 3-dimensional case. For small values of z the spheres do not intersect, while with growing z some overlap between the volumes starts to occur. At this point we are unable to offer a definitive strategy on how to tackle the integral (5.35), however we can outline a few possible alternatives on how to exploit the geometry just outlined:

1. for sufficiently large values of z it may be possible to approximate the union of the spheres $\{\mathcal{S}_k^{\mathbf{c},z}\}_{k=1}^{n!}$ with a sphere whose center is that of the sphere on which all the $\{\mathbf{c}_k\}_{k=1}^{n!}$ lay and radius equal to $R + z$. Referring to equations (5.30) and (5.31):

$$\bigcup_{k=1}^{n!} \mathcal{S}_k^{\mathbf{c},z} \approx \mathcal{S}^{\mathbf{C},R+z}$$

   This would make the calculation of the integral (5.35) easier since the region $\mathcal{W}(z)$ could be expressed as single sphere rather than $n!$;

2. the distribution being integrated is a high-dimensional Gaussian distribution. In high dimensions, it is well known that this distribution concentrates around a thin shell at a distance $\rho\sqrt{n}$ from the origin [142, 15]. Specific tail bounds might allow us to exploit this property and its interaction with the geometry $\mathcal{W}(z)$.

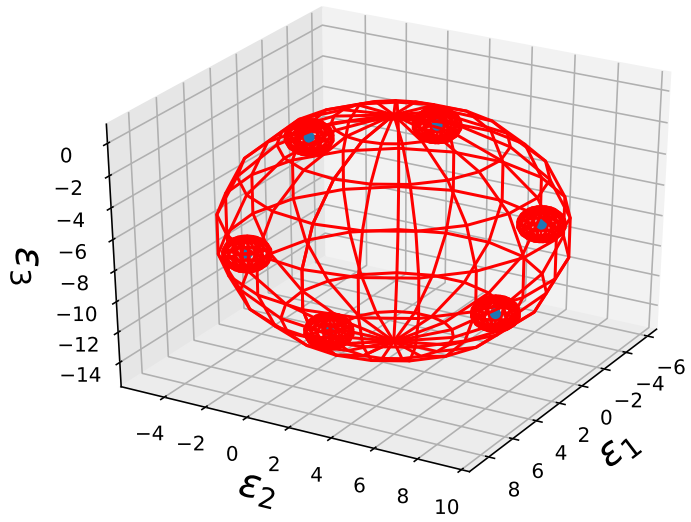


Figure 5-2: $\{\mathcal{S}_k^{\mathbf{c},z}\}_{k=1}^6$ and $\mathcal{S}^{\mathbf{C},R}$ based on a generic sine-wave type signal and $z = 1 < R = 7$.

5.3 Truncation of the inclusion-exclusion formula

So far we have explored possible approximations to the characterization of a TL-based likelihood through some theoretical arguments, without discussing approximations to the inclusion-exclusion formula. In order to complete the description, we now analyze the numerical behavior of possible truncations to it.

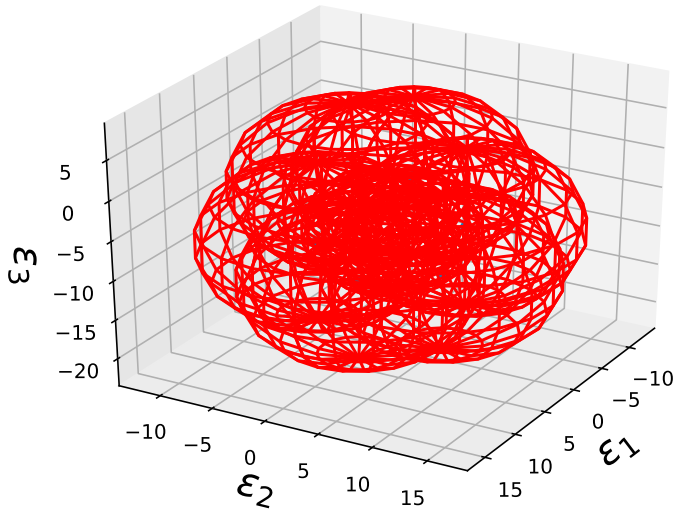


Figure 5-3: $\{\mathcal{S}_k^{\mathbf{c},z}\}_{k=1}^6$ and $\mathcal{S}^{\mathbf{C},R}$ based on a generic sine-wave type signal and $z = 10 > R = 7$.

5.3.1 The low-dimensional structure of the covariance $\mathbb{C}$ov(Z)

As already discussed in the previous section, the random nature of the Z_k variables is low dimensional. Therefore the covariance matrix must also reflect this property. We show this through some numerical experiments. We start by analyzing the covariance structure of the X variables. In Figure 5-5 we report the 9-by-9 matrix built by sampling a $n = 3$ signal with random values (uniformly samples between -10 and 10) and added Gaussian noise. As expected, the matrix is block diagonal as the X-s are only correlated if they correspond to the same $f(t_i)$ random variable and off-set by a different $g(t_j)$. The sample covariance values also coincide with the theoretically derived one as reported in Figures 5-4 and 5-6.

The question is now how this low-dimensional structure reflects on the Z random variables. We report the sample and analytical covariance matrix for the same $n = 3$ signal. In this case the covariance matrix is 6-by-6 and does not exhibit the clear block-structure of the $\mathbb{C}$ov(X) matrix. However it is still low-rank as the Z-s are generated

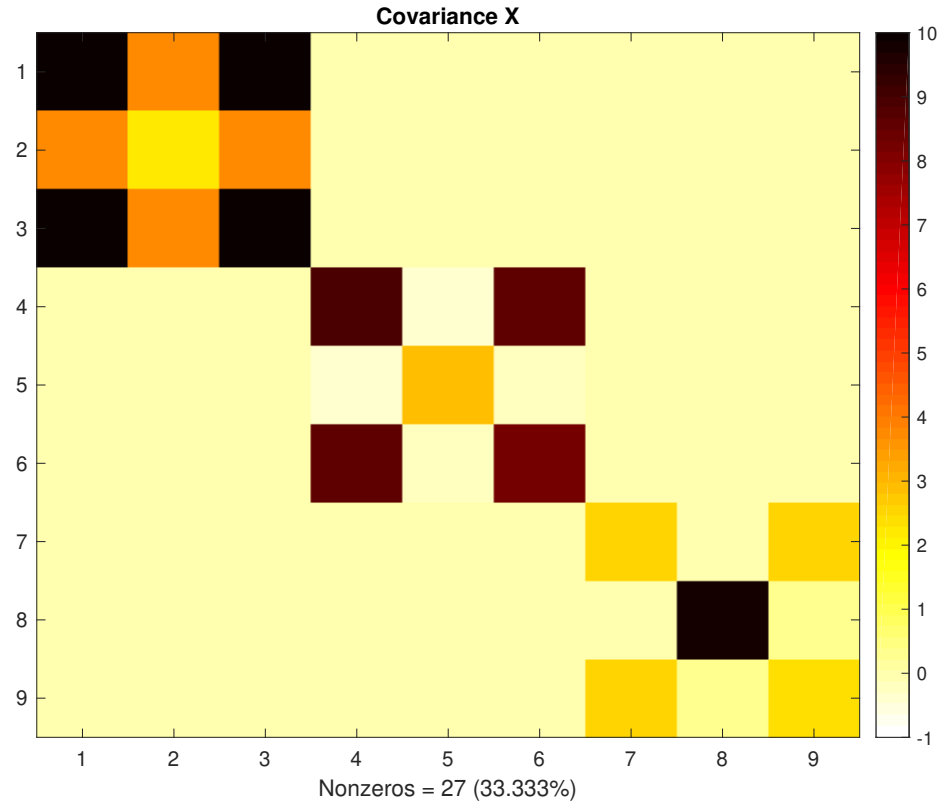


Figure 5-4: Analytical $\mathbb{C}$ov(X) for $n = 3$ sample signal.

through linear combinations of X-s. We verify this numerically and analytically. From a numerical standpoint the rank of the $\mathbb{C}$ov(Z) matrix is 3 and analytically it holds that [97]:

$$\text{rank}(\mathbb{C}\text{ov}(Z)) = \text{rank}(\rho^4 B \cdot \mathbb{C}\text{ov}(X) \cdot B^T) = \text{rank}(\mathbb{C}\text{ov}(X)).$$

We report in Figure 5-8 and 5-7 the sample and analytical $\mathbb{C}$ov(Z) matrices and their discrepancy (Figure 5-9). The main difference between the X and Z variables is that while in the first group of variables it is possible to identify subsets of random variable that are mutually independent, this is not possible for the Z random variables. This aspect seems to have a crucial impact on the degree of approximation that can be obtained by neglecting higher order terms in the summation (5.18).

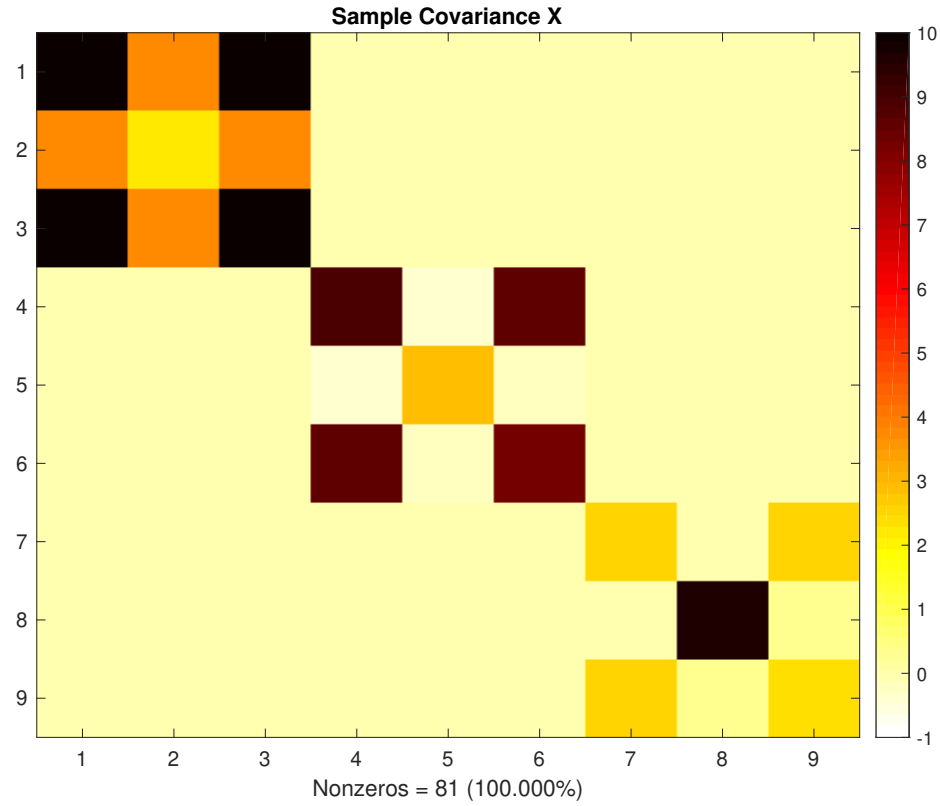


Figure 5-5: Sample $\mathbb{C}\text{ov}(\text{X})$ for $n = 3$ sample signal.

5.3.2 Approximating the inclusion-exclusion formula

For the case $n = 3$, we need to calculate:

$$\mathbb{P}(\min(Z) \leq z) = \text{CDF}(z) = \sum_{k=1}^{6} (-1)^{k-1} S_k$$

$$S_k = \sum_{\substack{\tau \in \mathcal{S}^{\{1,2,\ldots n!\}}_{\tau} \\ |\tau| = k}} \Phi_{\boldsymbol{\mu}_{\tau_1,\tau_2,\ldots,\tau_k}, \boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k}}(z) \tag{5.36}$$

which means each possible CDF for the subsets of cardinality up to 6 of the Z random variables. This means a total of 63 CDFs, i.e., terms in the summation. Unfortunately, as shown by the covariance matrices above, there is no specific ordering of the variables that would indicate a preferential way to drop terms in the summation.

Our approach is therefore that of analyzing the effect of successively adding terms to the inclusion/exclusion formula that refer to CDFs of subsets of increasing cardinality.

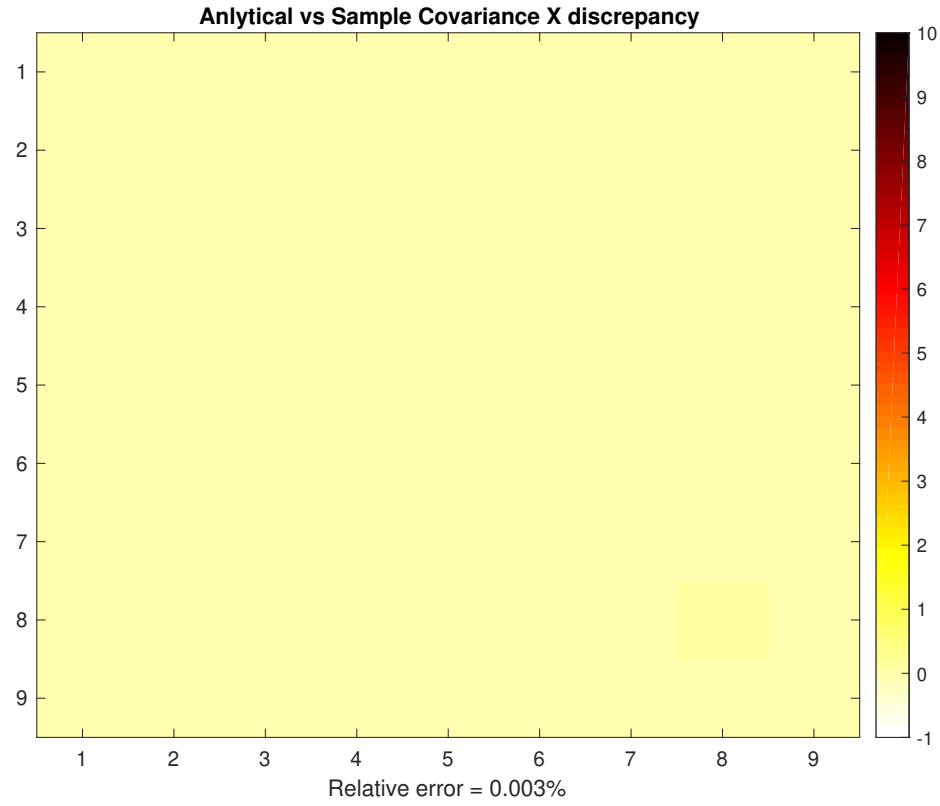


Figure 5-6: Absolute difference between sample and analytical $\mathbb{C}\text{ov}(\text{X})$ for $n = 3$ sample signal.

In Figure 5-10 we report the total summation of the CDFs in equation 5.18 up to cardinality k with $k = \{1, 2, 3, 4, 5, 6\}$. For a useful approximation, it could be considered satisfactory to drop all terms (i.e., k) that do not contribute significantly to the definition of the "curved" portion of the CDF. In Figure 5-10 it can be seen that starting with the 3rd or 4th order CDFs most of the monotonically increasing part of the function is defined. In other words, once $z \approx 200$ the CDF could be thresholded to 1 without the need to include the contribution of higher-order CDFs. In formulae:

$$\text{CDF}(z) \approx \begin{cases} \sum_{k=1}^{3} (-1)^{k-1} \sum_{\substack{\tau \in \mathcal{S}^{\{1,2,\ldots 6\}} \\ |\tau|=k}} \Phi_{\boldsymbol{\mu}_{\tau_1,\tau_2,\ldots,\tau_k}, \boldsymbol{\Sigma}_{\tau_1,\tau_2,\ldots,\tau_k}}(z) & \text{for } z \leq 200 \\ 1 & \text{for } z > 200 \end{cases} \tag{5.37}$$

The validity of this approximation holds even when applied to the PDF as shown in Figure 5-11. The validity of such approximation may depend on specific instantiation

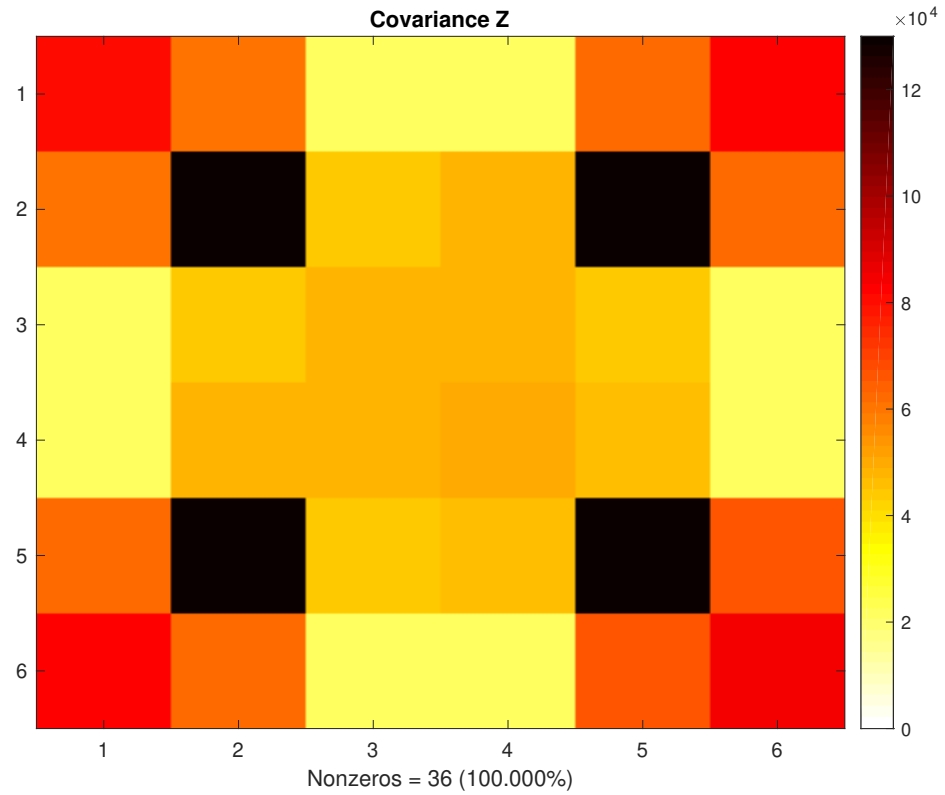


Figure 5-7: Analytical $\mathbb{C}\text{ov}(\text{Z})$ for $n = 3$ sample signal.

of the signals being analyzed and it is difficult to draw general conclusions. However, it can be observed, at least empirically, that the validity of such approximations largely depends on the structure of the covariance matrix. In Figures 5-12 and 5-13 we report the CDF built from a randomly generated covariance matrix that has full rank for a randomly generated 6-dimensional vector of random variables (i.e., uniformly drawn between –10 and 10). It clearly appears that the thresholding mechanism would work more effectively in this case than in the TL-specific case described above i.e., order $k = 3$ terms could be disregarded without affecting the accuracy of the approximation. The same observation holds if for the same vector a rank 3 covariance matrix was randomly generated (Figures 5-14 and 5-15). The rank-deficiency of the matrix does not seem to be the determining factor in the number of terms needed for a good approximation. We therefore conclude this subsection with a number of observations:

- based on a limited set of empirical tests it seems possible to neglect some number of terms in the identified expression for a TL-based CDF (equation (5.18));

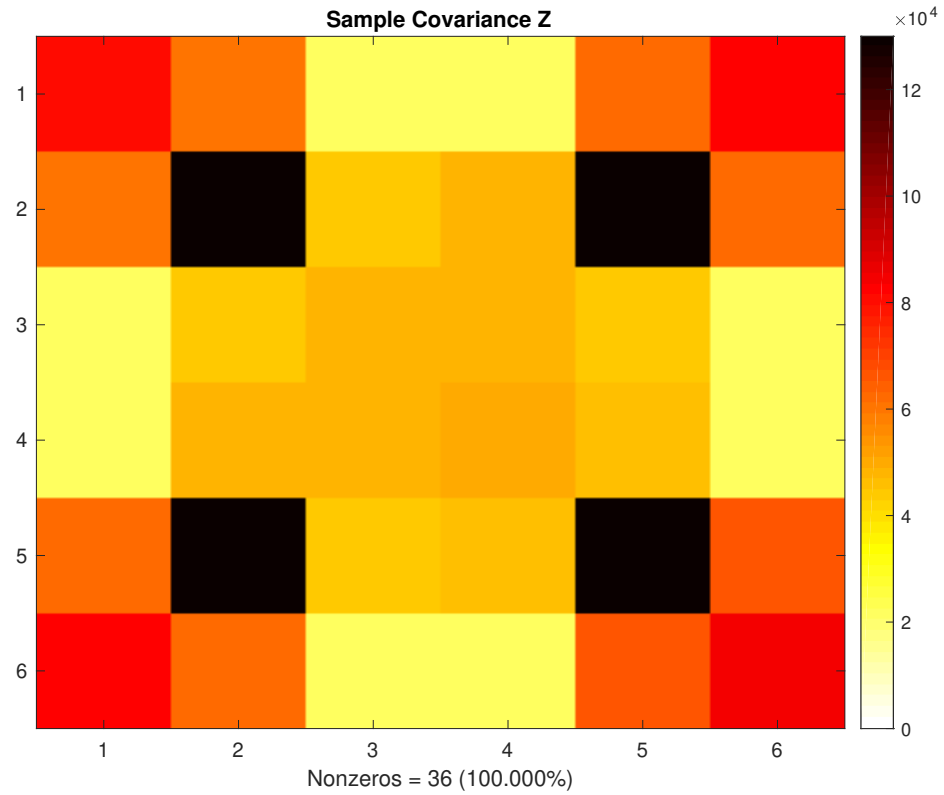


Figure 5-8: Sample $\mathbb{C}\mathrm{ov}(\mathrm{Z})$ for $n = 3$ sample signal.

- it appears that high-order (i.e., $k \geq 4$) joint CDFs (or PDFs) can be safely neglected through a thresholding-type mechanism;

- the structure of the covariance matrix, and the ordering of the random variables in the way the joint CDFs/PDFs are added to the summation defined in (5.18), seem to have an impact on the quality of the approximation that can be achieved through thresholding;

- for realistic applications (where $n \geq 100$) a useful approximation would need to limit the cardinality k to 1 or 2, or the number of terms would immediately increase to an unmanageable level.

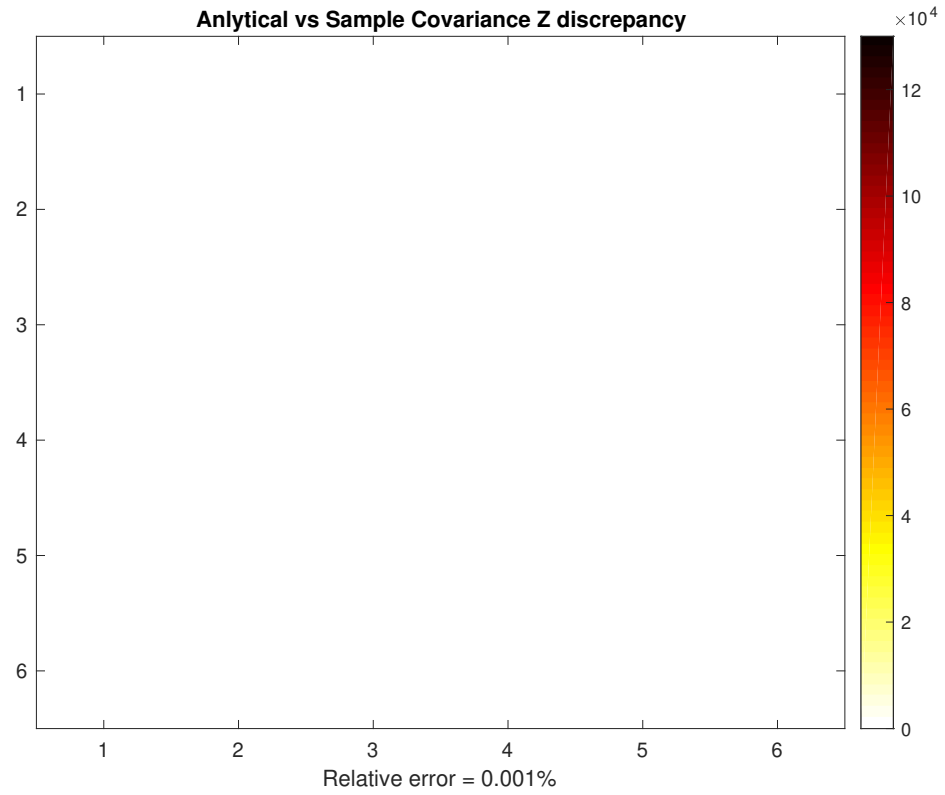


Figure 5-9: Absolute difference between sample and analytical $\mathbb{C}\text{ov}(\mathbf{Z})$ for $n = 3$ sample signal.

5.4 Empirical approximation

To conclude this chapter on the possible approximation to a TL-based likelihood function, we now proceed to characterize its behavior through simple Monte Carlo sampling.

Illustrative examples As a first exploration step we calculate the histogram of the TL distances between two signals $\mathbf{u}$ and $\mathbf{y}$ built as follows:

$$\mathbf{u} = A\left(\frac{1}{2}\sin(\omega\mathbf{t} + \varphi) + \frac{1}{2}\cos(4\omega\mathbf{t} + \varphi/4)\right) \tag{5.38}$$

$$\mathbf{y} = \mathbf{u} + \mathcal{N}(\mathbf{0}, \rho\mathbb{I}) \tag{5.39}$$

for 9 different pairs of ω and φ and $\rho = 1$, $A = 10$ (see Figure 5-16). Note in this specific case the signals are well-specified in that the only difference between them

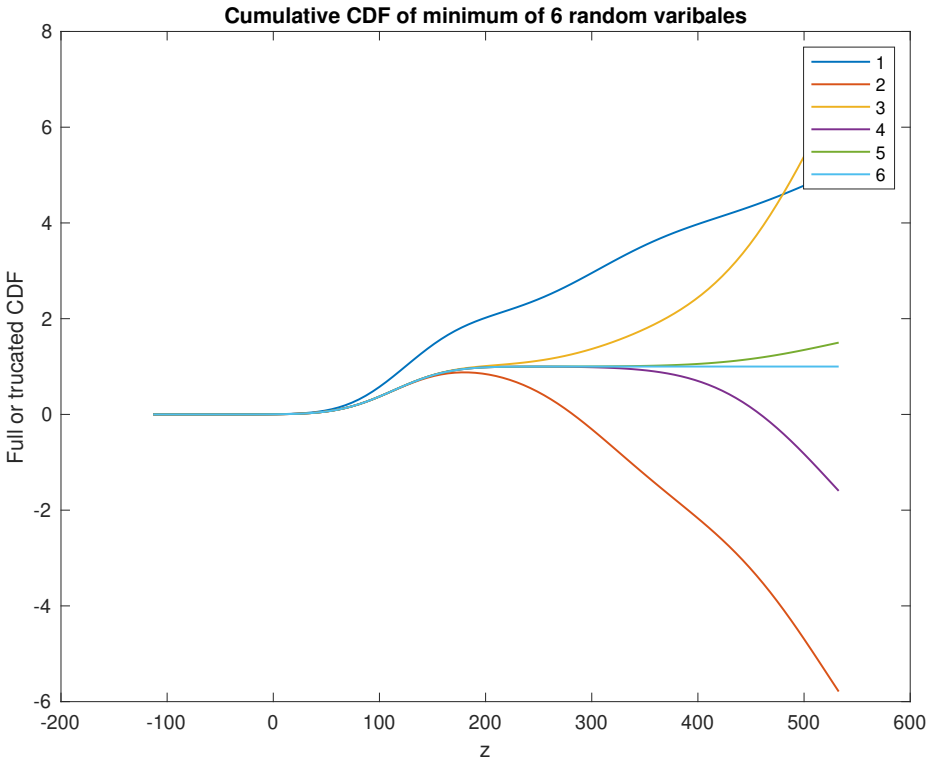


Figure 5-10: Truncated CDF of $\min(Z)$ for increasing cardinality orders k. For $k = 6$ the CDF reported is the exact one.

is the amount of noise. As such we are examining a special case of the more general discussion held in Section 5.1. We decide to also fit a Gamma distribution to this histogram and observe its proximity to Gaussianity according to the value taken by the shape parameter (α). We recall that a Gamma distribution is defined by two parameters: shape (α) and rate (β) parameters. Both parameters contribute to the definition of the means and variance of the distribution, however, the value of the shape parameter is what determines the asymptotically Gaussian behavior. In particular, the Gamma distribution can be viewed as a sum of exponential distributions for positive integer values of α: the sum of k exponential distributions with parameter β equals a Gamma distribution with parameter $\alpha = k$ and rate parameter equal to β [33]. By means of the central limit theorem, when k is high (i.e., $k > 30 - 50$), the Gamma distribution approaches the Gaussian one. In Figure 5-17 we observed a good fit of the gamma distribution in all 9 pairs of ω and φ, for an average α value of 60. This value indicates an almost Gaussian behavior, although not completely. With this level

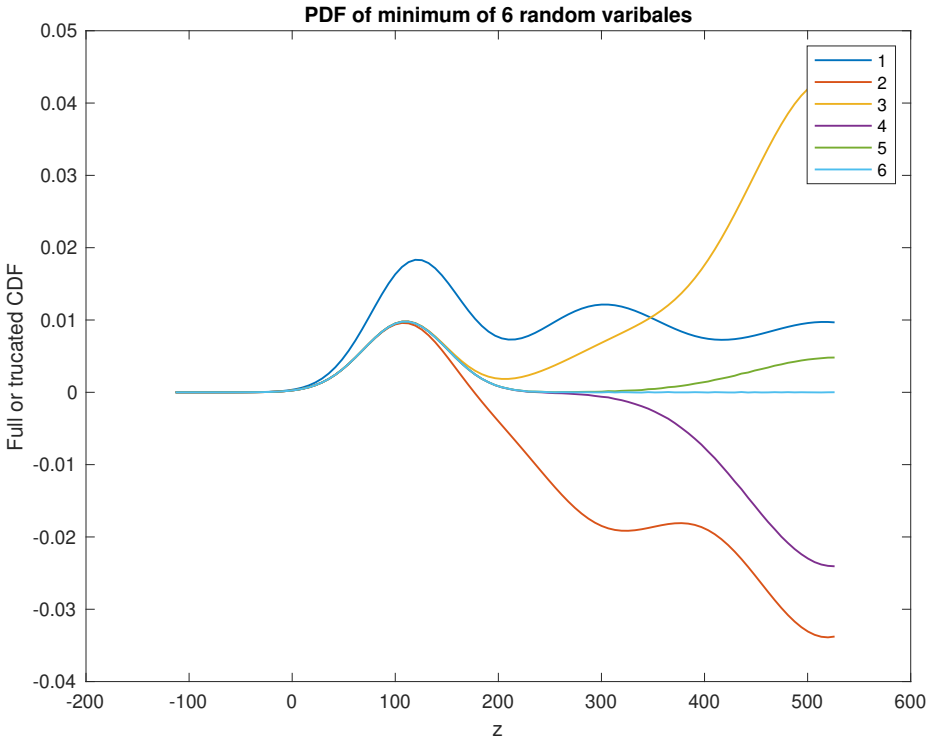


Figure 5-11: Truncated PDF of $\min(Z)$ for increasing cardinality orders k. For $k = 6$ the PDF reported is the exact one.

of noise and A we observe that the TL does not, on average, revert to the identity assignment when comparing the two signals (see Figure 5-18).

We therefore proceed with a more systematic study of how the Gamma fit and how it is impacted by changes in the noise level ρ and in particular on whether the Gaussian behavior is related to specific patterns in the mean assignment matrix. We report in Figure 5-19 a systematic mapping of the values of the shape parameter α as a function of ρ for the signals $\mathbf{u}$ and $\mathbf{y}$ as outlined above. It can be observed that the predominant behavior of the distribution is Gaussian for all regimes in which the noise level is either much larger or much smaller than the scale of the signal amplitude. All graphs in fact exhibit a "dip" around values of $\rho = 10^0$. This behavior is justifiable in statistical terms in that for low noise levels, the distance value is simply a sum randomly drawn i.i.d. chi-squared random variables, for which the central limit theorem applies. In a similar fashion, when the noise level is much higher than the signal scale, then the chi-squared random variable become *approximately* i.i.d and

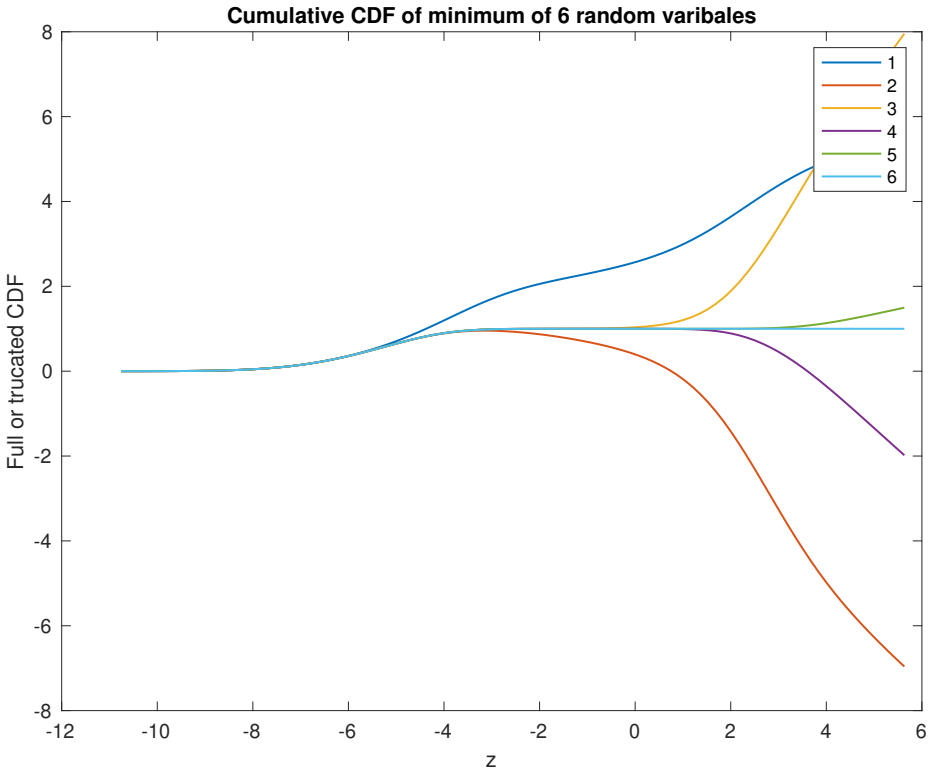


Figure 5-12: Comparison case: randomly generated 6-dimensional vector and randomly generated full-rank covariance matrix - Truncated CDF of $\min(Z)$ for increasing cardinality orders k. For $k = 6$ the CDF reported is the exact one.

the central limit theorem holds again. We conclude by repeating the experiment in a misspecified setting, i.e., where $\mathbf{u}$ and $\mathbf{y}$ are misspecified, i.e., even with minimal amount of noise, their difference will be non-zero. In this case, for lower values of noise the fitting is not meaningful since the optimal assignment will only be determined by the deterministic differences between the two signals and therefore be constant regardless of the noise draw (as such we have mostly omitted the values of fitted-α). For values of $\rho \geq 10^2$ the Gaussian approximation holds similarly to the case of a well-specified pair of signals given the high values of noise. For values of ρ between 1and10 the Gaussian approximation seems to be weaker for a larger portion of ρ-s than in the well specified case. This behaviour can be understood by the fact that the misspecification increases the differences in mean of the chi-squared random variables and invalidates the i.i.d. assumption until higher values of ρ (Figure 5-20).

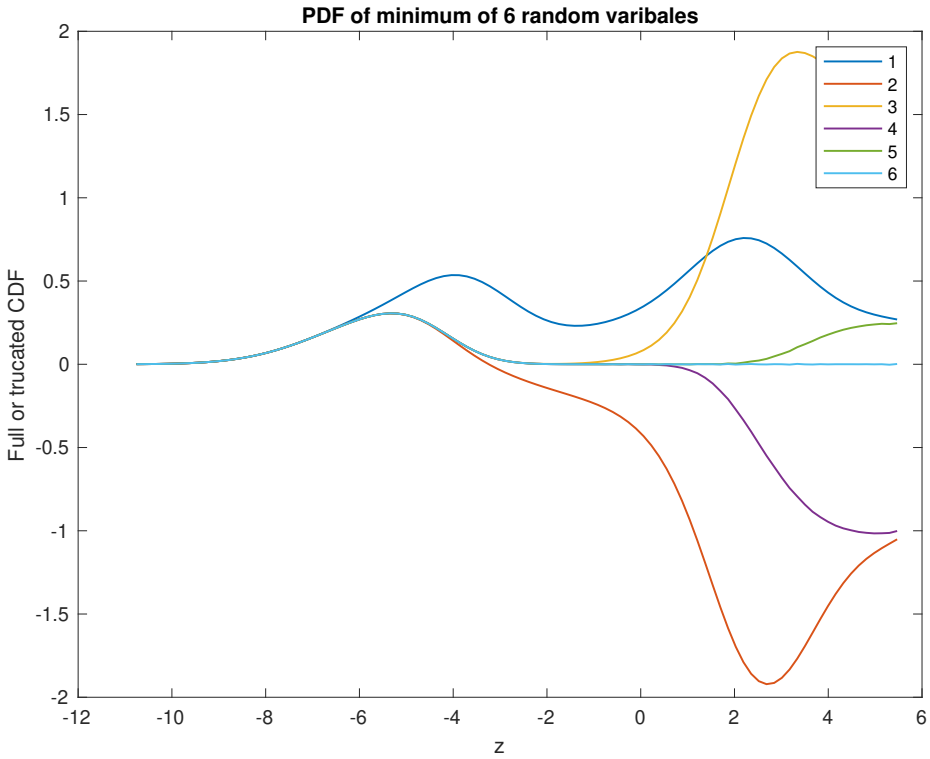


Figure 5-13: Comparison case: randomly generated 6-dimensional vector and randomly generated full-rank covariance matrix - Truncated PDF of $\min(Z)$ for increasing cardinality orders k. For $k = 6$ the PDF reported is the exact one.

5.5 Conclusions

In this chapter we have described the problem of the determining the nature of a TL-based likelihood function from three main viewpoints: analytical/exact, geometrically/approximate and numerically/approximate. While in Section 5.1 we were able to identify a close form expression for the large n behavior of the likelihood, this expression is intractable and therefore different strategies for its exploitation have been explored. Both the geometric and empirical approximations show some promising paths to research further, with the MCMC sampling in particular, showing a behavior of the empirical distribution easily associable to the well known Gamma family of distributions.

The discussions and ideas presented in the last two sections need further exploration to consolidate the intuitions into applicable results.

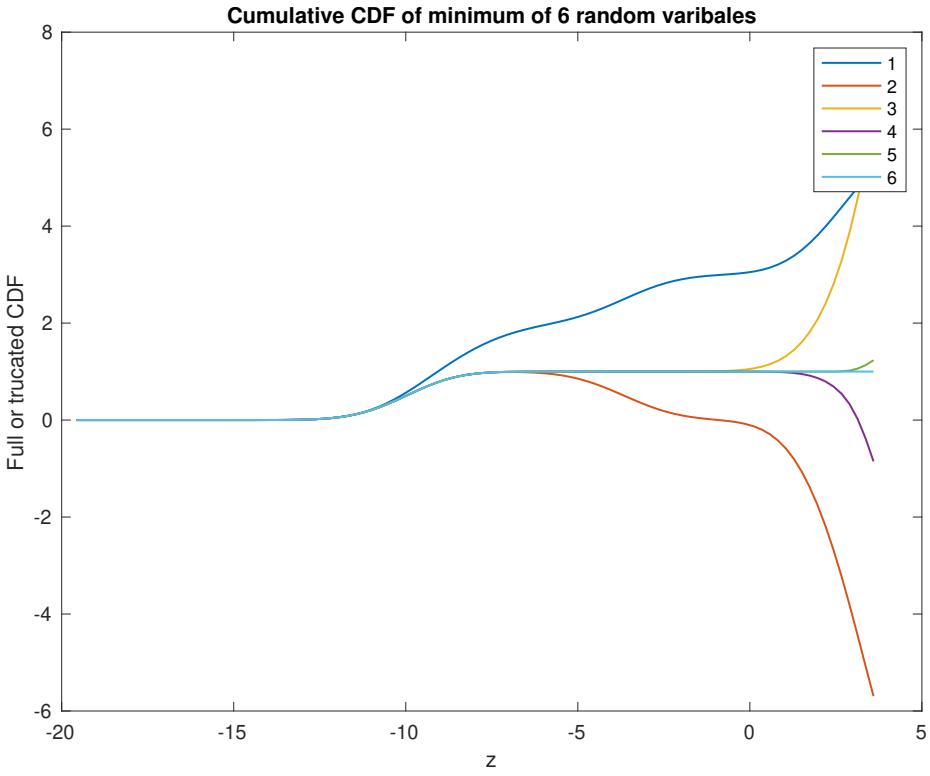


Figure 5-14: Comparison case: randomly generated 6-dimensional vector and randomly generated low-rank (rank 3) covariance matrix - Truncated CDF of $\min(Z)$ for increasing cardinality orders k. For $k = 6$ the CDF reported is the exact one.

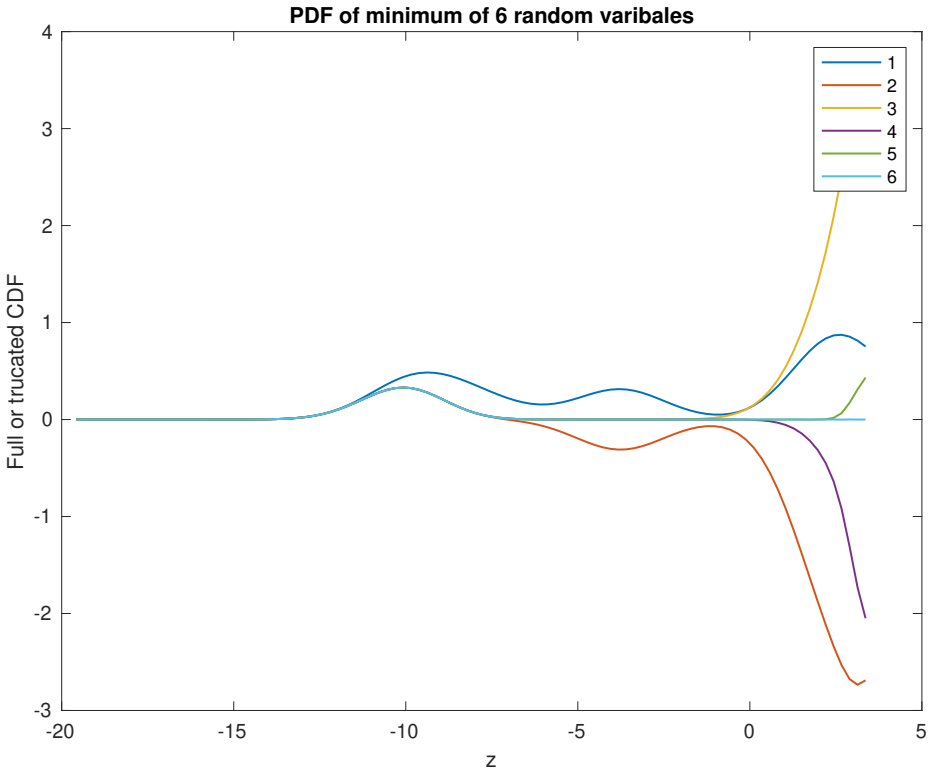


Figure 5-15: Comparison case: randomly generated 6-dimensional vector and randomly generated low-rank (rank 3) covariance matrix - Truncated PDF of $\min(Z)$ for increasing cardinality orders k. For $k = 6$ the PDF reported is the exact one.

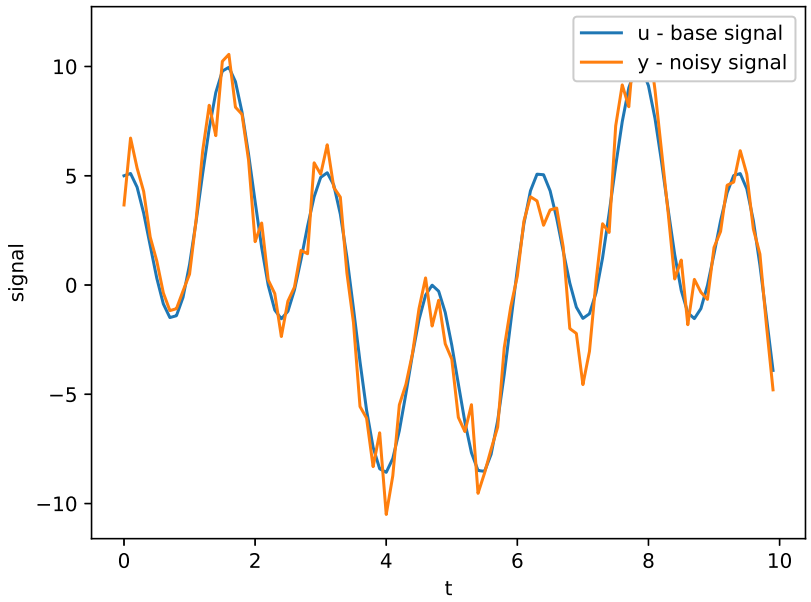


Figure 5-16: Sample $\mathbf{u}$ and $\mathbf{y}$.

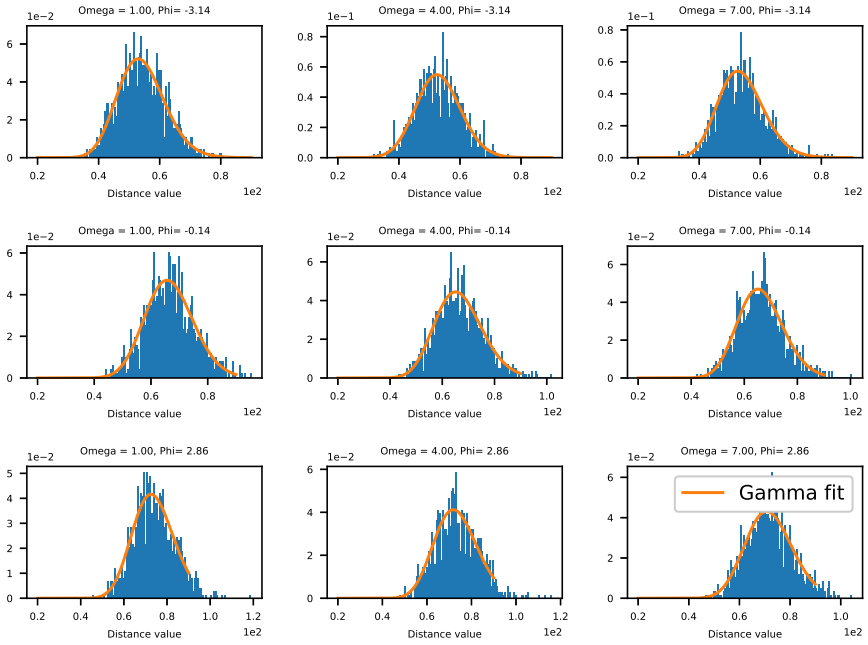


Figure 5-17: Histograms and Gamma fit for a $n_{samples} = 1000$ of TL$(\mathbf{u}, \mathbf{y})$ for $\rho = 1$. Average shape parameter $\alpha = 60$)

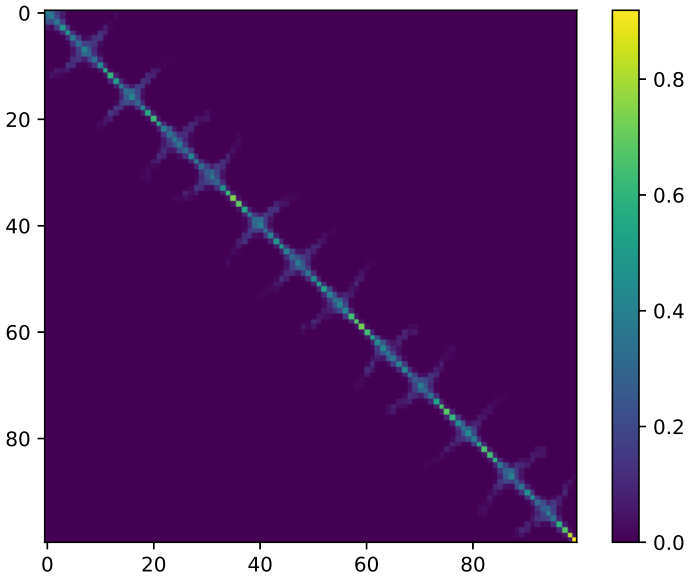


Figure 5-18: Average assignment matrix for the experiments shown in Figure 5-17.

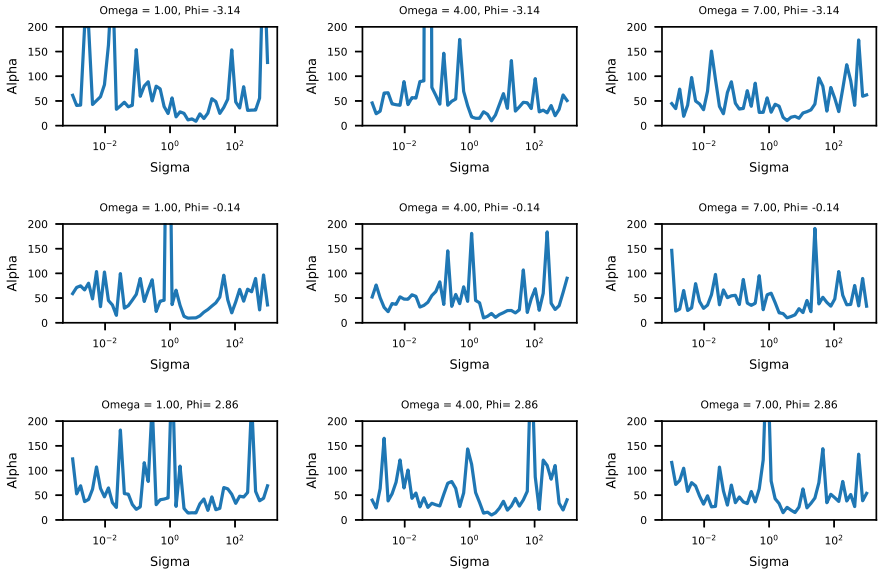


Figure 5-19: Value of α shape parameter for a Gamma fit for a $n_{samples} = 1000$ of TL$(\mathbf{u}, \mathbf{y})$ for varying values of noise (ρ).

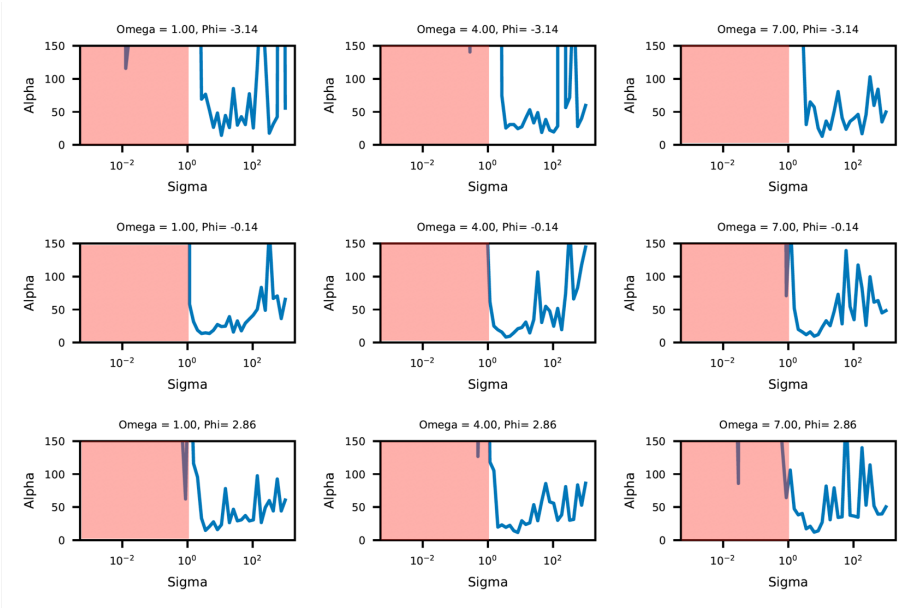


Figure 5-20: Value of α shape parameter for a Gamma fit for a $n_{samples} = 1000$ of TL$(\mathbf{u}, \mathbf{y})$ for varying values of noise (ρ) - Misspecified case.

Chapter 6

Optimal transport based linear regression

Linear regression is an essential problem across the natural sciences, social sciences, and engineering. The majority of linear regression approaches use a squared-error loss function, perhaps augmented with some regularization term (e.g., ridge regression [63] or the LASSO [131]). We formulate a linear regression problem that instead uses an optimal transport loss function, specifically the transport-Lagrangian (TL) distance. As already described earlier in this thesis, the TL distance is well suited to signal and image analysis, as it allows for the comparison of responses across *different* values of the explanatory variables of the regression model—for instance, time or space coordinates. Computationally, however, regression with the TL distance leads to a challenging optimization problem: the search space is not only that of the coefficients for the linear combination (typically $\mathbb{R}^m$), but also that of all possible permutations $\sigma \in \mathrm{Perm}(n)$ of the data vector. The space of the permutations is by nature discrete and the classic notion of gradient cannot be applied directly. The general problem of linear regression with permuted data as an additional degree of freedom in the optimization has already been explored in a variety of contexts unrelated to signal processing [1, 85, 41]. In all of these applications, the challenges of the optimization problem ([100]) have been described and tackled in different ways. In some cases,

constraints on the sparsity of the associated permutation matrix are imposed, as well as on its distance from the identity matrix [86]. Some convex hull-relaxations have also been proposed [41]. The main differences between what we propose in this chapter and previous literature is:

- a new application context based on model misspecification and functional data;
- the introduction, via TL-distance, of a regularizing term - the horizontal cost - on the possible forms the permutation matrix can take , i.e., discouraging strong departures form the identity permutation.

In this chapter we provide a geometric description of the optimization problem and use this geometry to propose several minimization algorithms. We test these algorithms and demonstrate the usefulness of the TL regression in a variety of application problems.

6.1 Geometry and algorithms

Let us consider a linear model of the form:

$$\mathbf{u}(\mathbf{x}) = \mathbf{\Phi}(\mathbf{x})\boldsymbol{\beta} \tag{6.1}$$

where $\mathbf{x}$ represents some set of time-space coordinates, $\mathbf{\Phi}$ is a discrete forward model operator (e.g., set of basis functions) and $\boldsymbol{\beta}$ is the vector containing the coefficients to be estimated. Setting some notation: $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{\Phi}(\mathbf{x}) \in \mathbb{R}^{n\times m}$, $\boldsymbol{\beta} \in \mathbb{R}^m$ and $\mathbf{u}(\mathbf{x}) \in \mathbb{R}^n$. The classic least-squares estimation method for such linear problems is formulated as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} ||\mathbf{y} - \mathbf{\Phi}(\mathbf{x})\boldsymbol{\beta}||_2^2 \tag{6.2}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector containing the observed data. This problem has a closed form solution when $\mathbf{\Phi}(\mathbf{x})$ has full column rank. We now intend to use the transport-Lagrangian distance as a loss function. Given two vectors $\mathbf{y}, \mathbf{u}$, which we interpret as

discretized signals, the TL_p^λ distance, is defined as follows:

$$\min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^m \\ \boldsymbol{\sigma}\in\mathrm{Perm(N)}}} ||\mathbf{y}_{\boldsymbol{\sigma}} - \boldsymbol{\Phi}(\mathbf{x})\boldsymbol{\beta}||_p^p + \lambda||\mathbf{x}_{\boldsymbol{\sigma}} - \mathbf{x}||_p^p \tag{6.3}$$

where $\boldsymbol{\sigma}$ indicates any permutation of the elements of the vector $\mathbf{y}$, i.e., $\mathbf{y}_{\boldsymbol{\sigma}} = [y_{\sigma(1)}, \ldots, y_{\sigma(i)}, \ldots, y_{\sigma(n)}]$. The first term of the above expression resembles the one of any ℓ_p norm: amplitude comparison between $\mathbf{u}$ and $\mathbf{y}$ at a each coordinate point $\mathbf{x}$. The second one is meant to control the amount of across-coordinate transport induced by the optimization over $\boldsymbol{\sigma}$. The parameter λ is an additional degree of freedom to be chosen to control the relative weight of the two terms of the objective function. Since it will not affect the algorithmic analysis of this chapter, we will set it to $\lambda = 1$. We also focus on $p = 2$, i.e., the TL_2^1 distance, resulting in:

$$\min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^m \\ \boldsymbol{\sigma}\in\mathrm{Perm(n)}}} ||\mathbf{y}_{\boldsymbol{\sigma}} - \boldsymbol{\Phi}(\mathbf{x})\boldsymbol{\beta}||_2^2 + ||\mathbf{x}_{\boldsymbol{\sigma}} - \mathbf{x}||_2^2 \tag{6.4}$$

The challenge introduced by the transport-Lagrangian distance as a misfit measure is that the regression problem will have to be solved by optimizing not only over $\boldsymbol{\beta}$, but also over $\boldsymbol{\sigma}$. The set of all possible permutations of the elements $\mathbf{y}$ vector contains at most $n!$ distinct elements, which makes enumeration an impractical strategy even for relatively coarse discretizations (e.g., $N = 100$).

6.1.1 Computational strategies

Coordinate descent For a given permutation $\boldsymbol{\sigma}$, the least-squares optimization problem in the parameters $\boldsymbol{\beta}$ has a closed-form solution, given by:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\sigma}} = \left(\boldsymbol{\Phi}(\mathbf{x})^T\boldsymbol{\Phi}(\mathbf{x})\right)^{-1}\boldsymbol{\Phi}(\mathbf{x})^T\mathbf{y}_{\boldsymbol{\sigma}} \tag{6.5}$$

From a geometric viewpoint, $\hat{\boldsymbol{\beta}}$ represents the vertex (minimum) of a convex quadratic surface (elliptic paraboloid). Vice-versa, for a given value of the parameters $\boldsymbol{\beta}$, it

is possible to find the solution to the optimal assignment problem through a linear programming approach (specialized auction algorithm) with a worst case complexity of $O(n^3)$. This naturally suggests an alternating approach between finding the optimal assignment $\boldsymbol{\sigma_{opt}}$ and $\hat{\boldsymbol{\beta}}$ (in a coordinate-descent fashion) over the space of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. The challenge is that the convergence of such a approach to the global minimum is not guaranteed, since the problem is not convex in both $\boldsymbol{\sigma}$ and $\boldsymbol{\beta}$. In fact, the $n!$ vertices of the paraboloids defined by each of the permutations in the least squares problem over $\boldsymbol{\beta}$ are the vertices of a permutohedron that has $2^n - 2$ faces [105]. This implies there can be as many as $n!$ local minima that could cause a coordinate descent approach to fail. When the dimensionality m of $\boldsymbol{\beta}$ is sufficiently low, a grid search on region of $\mathbb{R}^m$ could be acceptable. However, this approach becomes impractical when m grows larger. Similarly, enumerating all the possible $n!$ permutations $\boldsymbol{\sigma}$ is impractical in the most common scenarios, where for example $n > 100$.

A geometric approach We propose instead an approach that exploits the geometry of the problem. Let us fix a permutation $\boldsymbol{\sigma}$. The equation of the associated paraboloid in $\boldsymbol{\beta}$ writes as follows:

$$Z_{\boldsymbol{\sigma}}(\boldsymbol{\beta}) = ||\mathbf{y}_{\boldsymbol{\sigma}} - \boldsymbol{\Phi}(\mathbf{x})\boldsymbol{\beta}||_2^2 + ||\mathbf{x}_{\boldsymbol{\sigma}} - \mathbf{x}||_2^2 \tag{6.6}$$

where the coordinates-dimensions in which the paraboloid lives are specified by the vector $\boldsymbol{\beta}$. Equation (6.5) gives us the expression for the minimum of a given paraboloid ($z_{\boldsymbol{\sigma}}$), which we substitute in (6.6):

$$Z(\mathbf{y}_{\boldsymbol{\sigma}}) = ||\mathbf{y}_{\boldsymbol{\sigma}} - \boldsymbol{\Phi}(\mathbf{x})\big(\boldsymbol{\Phi}(\mathbf{x})^T\boldsymbol{\Phi}(\mathbf{x})\big)^{-1}\boldsymbol{\Phi}(\mathbf{x})^T\mathbf{y}_{\boldsymbol{\sigma}}||_2^2 + ||\mathbf{x}_{\boldsymbol{\sigma}} - \mathbf{x}||_2^2 \tag{6.7}$$

Let us call $\mathbf{A} = \boldsymbol{\Phi}(\mathbf{x})\big(\boldsymbol{\Phi}(\mathbf{x})^T\boldsymbol{\Phi}(\mathbf{x})\big)^{-1}\boldsymbol{\Phi}(\mathbf{x})^T$ and rewrite:

$$\begin{aligned} Z(\mathbf{y}_{\boldsymbol{\sigma}}) &= ||(\mathbb{I} - \mathbf{A})\mathbf{y}_{\boldsymbol{\sigma}}||_2^2 + ||\mathbf{x}_{\boldsymbol{\sigma}} - \mathbf{x}||_2^2 && (6.8)\\ &= \mathbf{y}_{\boldsymbol{\sigma}}^T(\mathbb{I} - \mathbf{A})^T(\mathbb{I} - \mathbf{A})\mathbf{y}_{\boldsymbol{\sigma}} + 2||\mathbf{x}||_2^2 - 2\mathbf{x}^T\mathbf{x}_{\boldsymbol{\sigma}} && (6.9)\end{aligned}$$

By temporarily lifting the restriction on the discrete nature of the variable $\boldsymbol{\sigma}$, and ignoring the horizontal transport term $\mathbf{x}^T\mathbf{x}_{\boldsymbol{\sigma}}$, we can read equation (6.7) as that of a paraboloid in the variable-coordinates $\boldsymbol{y} \in \mathbb{R}^N$. This paraboloid, centered at the origin, must contain on its surface all of the $\boldsymbol{y}_{\boldsymbol{\sigma}}$ points that are defined by all possible permutations of the data vector $\mathbf{y}$. Reintroducing the horizontal term, we will simply be specifying a different $z-$axis intercept of the aforementioned paraboloid for each $\boldsymbol{\sigma}$. The permutation $\boldsymbol{\sigma}_{opt}$ that would solve the original problem (6.4) can be geometrically identified as the one that sets the point $\boldsymbol{y}_{\boldsymbol{\sigma}}$ the closest to the origin. The problem can be equivalently formulated as follows:

$$\min_{\boldsymbol{\sigma}} \ \mathbf{y}_{\boldsymbol{\sigma}}^T\mathbf{Q}\mathbf{y}_{\boldsymbol{\sigma}} - 2\mathbf{x}^T\mathbf{x}_{\boldsymbol{\sigma}}, \tag{6.10}$$

where $\mathbf{Q} = (\mathbb{I} - \mathbf{A})^T(\mathbb{I} - \mathbf{A})$ is symmetric positive semi-definite. This problem is an *integer semi-definite quadratic optimization* problem, which can be solved through specific techniques such as branch and bound, pre-solve, cutting-planes, etc. [151]. While these problems are generally NP-hard, they still represent a more systematic and efficient way of solving the problem, rather than a brute-force algorithm (implying the enumeration of all possible permutations $\boldsymbol{\sigma}$) or alternating between optimizing over $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ as discussed above.

Simplifications to the computational solution of this problem can come both from continuous relaxations as well as exclusions of the horizontal transport term. Neglecting the horizontal term may have varying practical implications in terms of approximations of the original solution. Same applies to continuous relaxations of the problem. Under certain conditions of positive semi-definiteness of the quadratic form (i.e., $\mathbf{Q}$) and convexity of the feasible set, the relaxed solution may coincide with that of the integer problem. In the next subsections we will not explore these options, but rather use a numerical mixed-integer solver to test the advantage of this formulation in a misspecified setting versus the classic regression setup.

6.2 Numerical experiments

In this section we test the above formulation of the problem against a number of synthetic and then more realistic applications. In all of the cases the minimization problem has been solved by using GUROBI integer-programming optimizer [61].

6.2.1 Synthetic warping

Data generation and inference model In this set of experiments we try to recover the coefficients of a linear combination of some basis functions. In particular we generate our data-set through:

$$\begin{aligned}\mathbf{y} \sim \quad & a_1 \exp(-0.2\mathbf{t})\cos(2\mathbf{t}) + b_1 \exp(-0.1\mathbf{t})\sin(2\mathbf{t}) \\ & + a_2 \exp(-0.1\mathbf{t})\cos(5\mathbf{t}) + b_2 \exp(-0.4\mathbf{t})\sin(5\mathbf{t}) \\ & + a_3 \cos(6\mathbf{t}) + b_3 \sin(6\mathbf{t}) + \mathcal{N}(0, \mathbb{I}) \text{ with } \mathbf{t} \in [0, 10]\end{aligned}$$

with:

$$\begin{aligned} a_1 &= 10 & b_1 &= 8 \\ a_2 &= 2 & b_2 &= 3 \\ a_3 &= 6 & b_3 &= 5 \end{aligned}$$

We report a realization of $\mathbf{y}$ in Figure 6-1 with $n = 100$ discretization points. The objective will be to recover these coefficients (linear regression), with a model, however, that exhibits a non-identity warping around the vector of time indices , i.e., $\mathbf{t}^* = h(\mathbf{t})$.

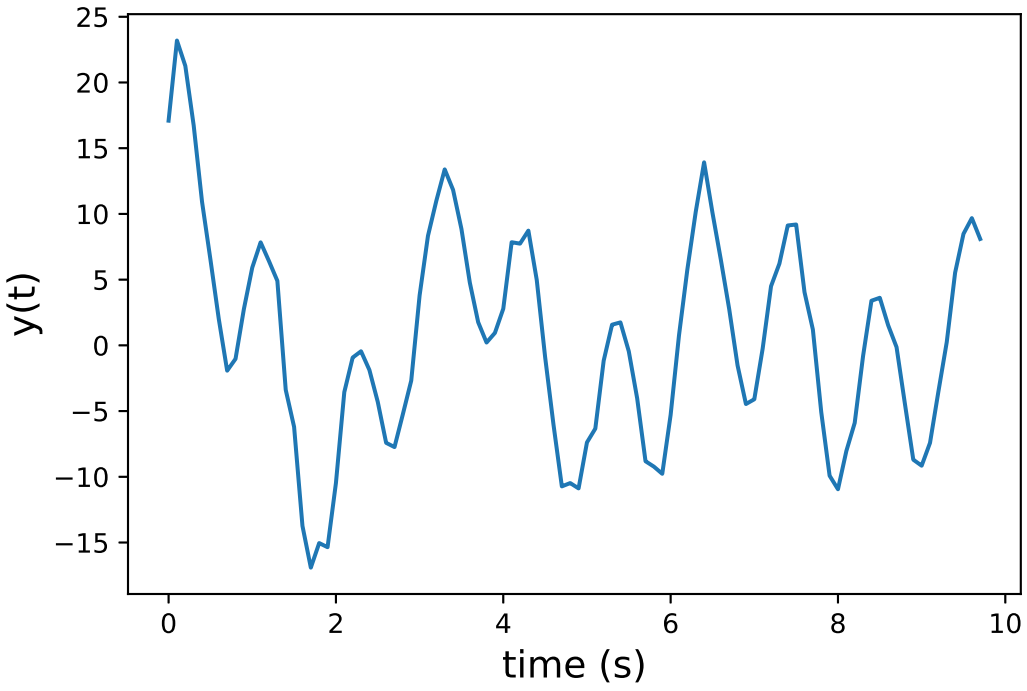


Figure 6-1: Sample data vector

In particular:

$$\begin{aligned}\mathbf{u}^* \sim & a_1 \exp(-0.2\mathbf{t}^*)\cos(2\mathbf{t}^*) + b_1 \exp(-0.1\mathbf{t}^*)\sin(2\mathbf{t}^*) \\ & + a_2 \exp(-0.1\mathbf{t}^*)\cos(5\mathbf{t}^*) + b_2 \exp(-0.4\mathbf{t}^*)\sin(5\mathbf{t}^*) \\ & + a_3 \cos(6\mathbf{t}^*) + b_3 \sin(6\mathbf{t}^*) + \mathcal{N}(0, \mathbb{I})\end{aligned}$$

Warping definition ($h(\mathbf{t})$) We generate the warping through the following expression:

$$h(\mathbf{t}) = C_0 + C_1 \int_0^{\mathbf{t}} \exp\left(W(z)\right) dz \tag{6.11}$$

where the the constants C_0 and C_1 are chosen such that $h(0) = 0$ and $h(10) = 10$ (i.e., the first and last point of the data time vector are not warped), and the exponentiation of $W(\mathbf{t})$ ensure the monotonicity of the mapping (see [106]). The $W(\mathbf{t})$ are instead

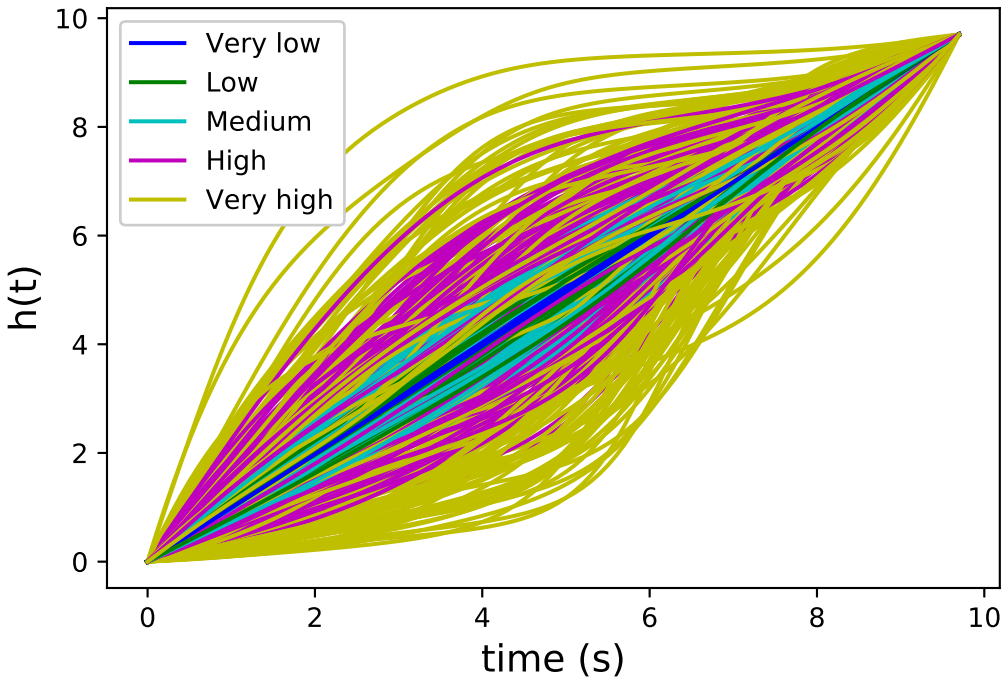


Figure 6-2: Samples of warping function corresponding to five different levels of intensity $A = \{0.01, 0.1, 0.25, 0.5, 1\}$

samples from a Gaussian process $\mathcal{GP}$ of zero-mean and squared exponential covariance kernel $\mathbf{k}(t, t') = \exp(-(t - t')^2/(2l^2))$ with $l = 2$. To simulate this kind of process we used a closed form expression for the eigenvalue/eigenfunction Karhunen-Loeve expansion, in formulae:

$$\mathcal{GP}(\mathbf{0}, \mathbf{k}(t, t')) \sim A \cdot \sum_{j=1}^{J} \lambda_j \phi_j(\mathbf{t}) \zeta_j \text{ where } \zeta_j \sim \mathcal{N}(0, 1) \tag{6.12}$$

where the eigenvalues λ_j and eigenfunctions ϕ_j (Hermite polynomials) are as detailed in [150]. The parameter A is instead introduced by us to calibrate the *intensity* of the warping , i.e., the level of misspecification. We report in Figure 6-2 a plot of some samples of $h(\mathbf{t}^*)$ for five different levels (i.e., A) of intensity.

Results For each of the intensity levels we sample $n_{samples} = 500$ warpings $h(\mathbf{t}^*)$ and generate the associated models. We then invert the data $\mathbf{y}$ for each of these models

Levels	a_1	a_2	b_1	b_2	c_1	c_2	Average
Very Low (A = 0.01)	-0.1	-0.01	-0.06	-0.28	-0.16	-0.07	-0.11
Low (A = 0.1)	0.84	0.59	0.67	0.18	0.84	1.36	0.75
Medium (A = 0.25)	0.19	1.54	0.76	-0.95	1.71	1.99	0.88
High (A = 0.5)	-0.78	1.64	0.28	-3.98	1.3	1.75	0.04
Very High (A = 1)	-2.48	0.67	-0.89	-9.36	0.53	0.5	-1.84

Table 6.1: $\ell_{2_{\text{RMSE}}} - \text{TL}_{2_{\text{RMSE}}}$ over $n_{samples}$ for each model coefficient and warping intensity level. Last column contains an average across the coefficients. Color-coding: red, if $\ell_{2_{\text{RMSE}}} < \text{TL}_{2_{\text{RMSE}}}$, green if $\ell_{2_{\text{RMSE}}} > \text{TL}_{2_{\text{RMSE}}}$

.

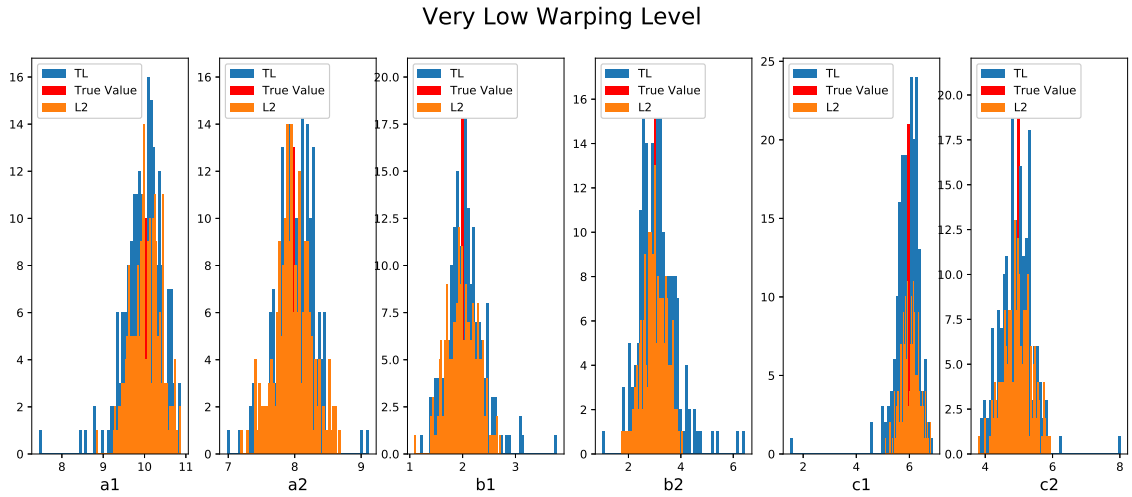


Figure 6-3: Histogram of recovered coefficients for $n_{samples} = 500$ of $A = \{0.01\}$ (very low level) warping $h(\mathbf{y})$

using both the TL and ℓ_2-based regression. We report in Figures 6-3, 6-4, 6-5, 6-6, 6-7 the histograms of the recovered coefficients for each sampled model , i.e., the result of the inversion of the data $\mathbf{y}$ with each sampled model $\mathbf{u}(h(\mathbf{t}))$.

Analysis In order to quantitatively interpret the results as shown in the histograms, we first calculate the difference in RMSE (root mean square error) between the ℓ_2 and TL recovered coefficients. We do so for each intensity level. The values are reported in table 6.1. We highlighted in red the scores for which the TL performed worse than the ℓ_2 , i.e., the corresponding RMSE is lower for the ℓ_2 than the TL. Vice versa, when the squared error was higher for the ℓ_2 than the TL, we plotted the corresponding

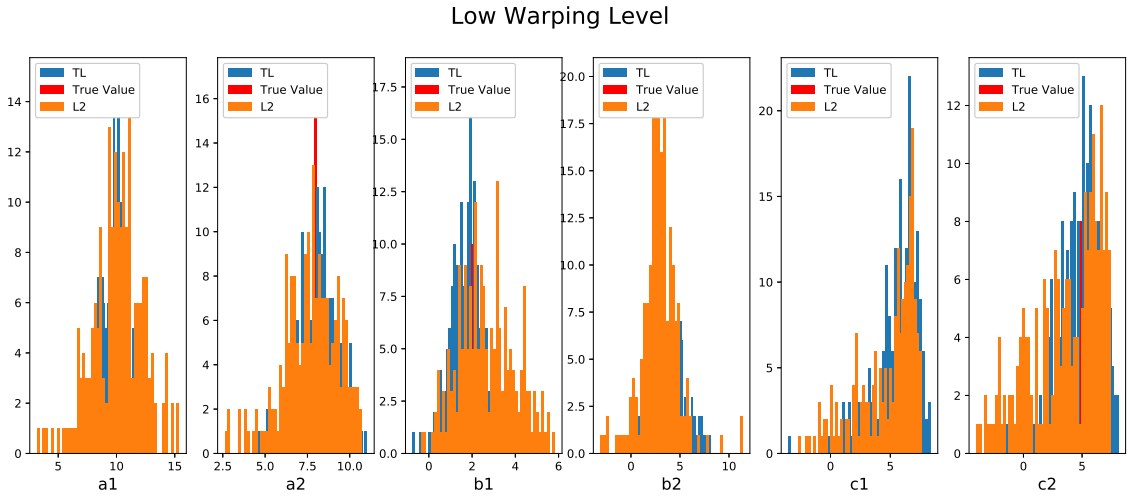


Figure 6-4: Histogram of recovered coefficients for $n_{samples} = 500$ of $A = \{0.1\}$ (low level) warping $h(\mathbf{y})$

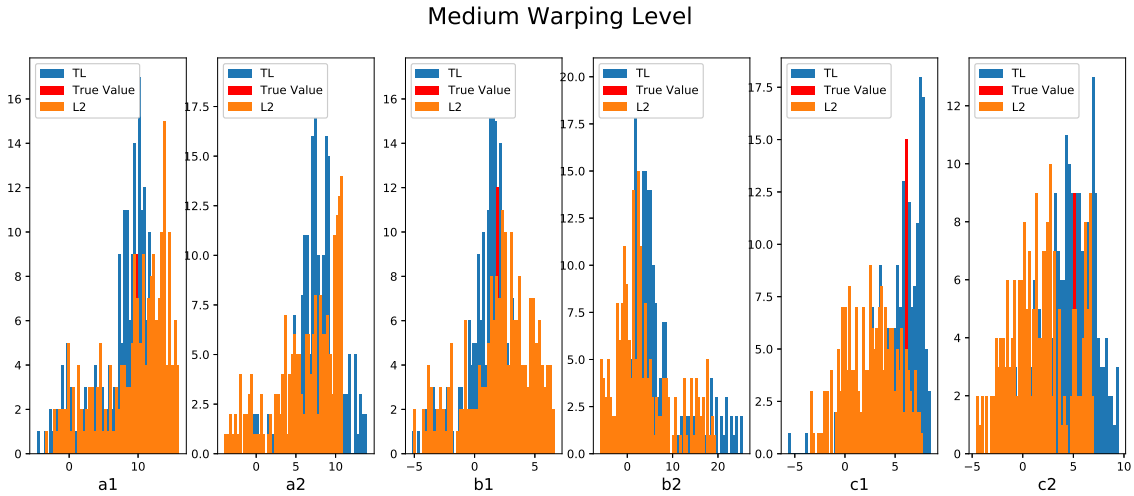


Figure 6-5: Histogram of recovered coefficients for $n_{samples} = 500$ of $A = \{0.25\}$ (medium level) warping $h(\mathbf{y})$

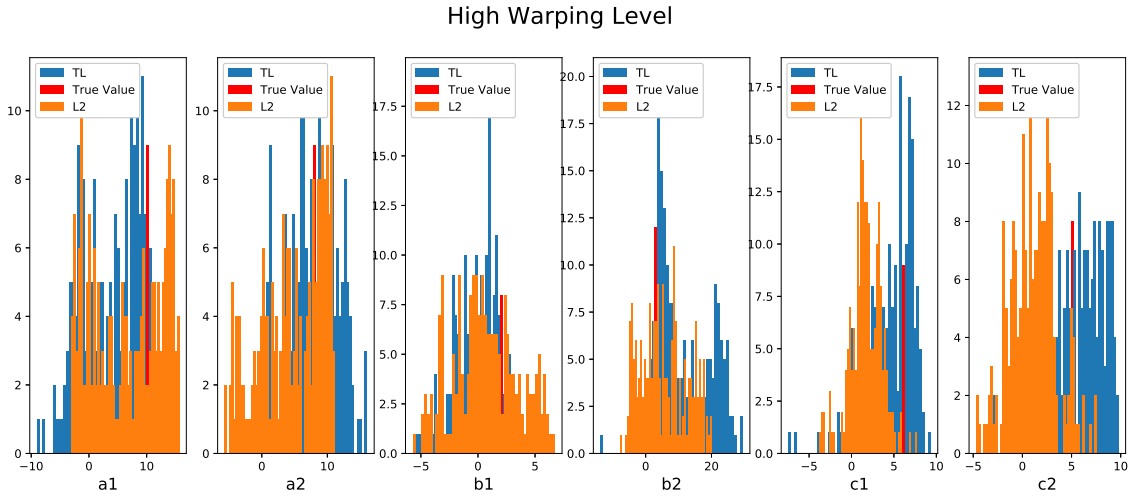


Figure 6-6: Histogram of recovered coefficients for $n_{samples} = 500$ of $A = \{0.5\}$ (high level) warping $h(\mathbf{y})$

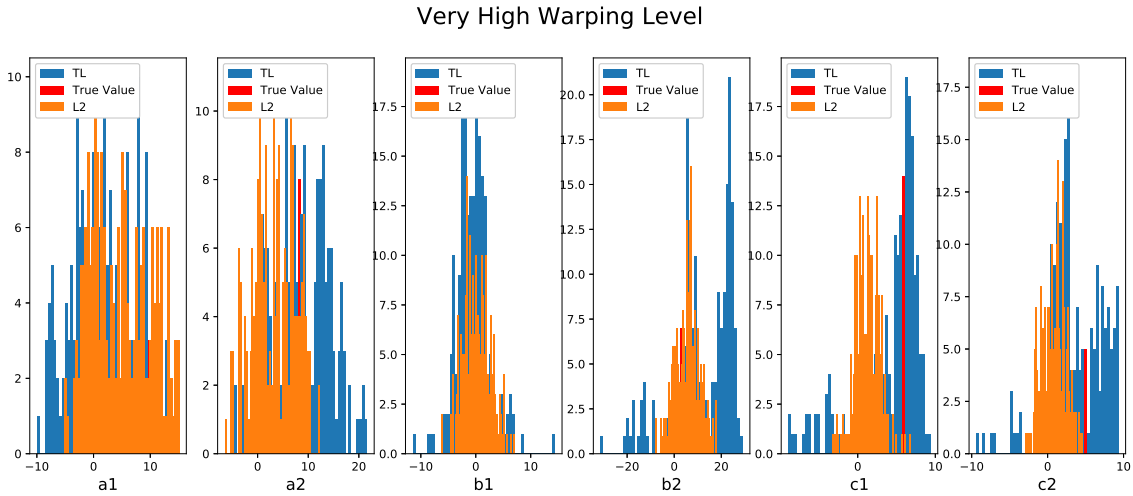


Figure 6-7: Histogram of recovered coefficients for $n_{samples} = 500$ of $A = \{1\}$ (very high level) warping $h(\mathbf{y})$

difference in score in green. To facilitate the interpretation of these results further, we also report the box plot of the absolute errors for each intensity level, averaged across the signal coefficients in Figure 6-8. The main emerging trend is that the TL lowers the average absolute error in the recovery of the model coefficients when the warping intensity is neither too low or too high (see Figure 6-9). In the former case, the performance of the two distances is in fact equivalent, as expected. In the latter case, the amount of misspecification is high enough for the TL to transport enough mass to induce bi-modal histograms. This is particularly evident for coefficients a_2 and b_2 in Figure 6-7, where the TL distance recovers two main values as possible "best fits." In this sense, the higher values of RMSE for this warping intensity are mainly due to the bi-modality of these histograms rather than the TL recovering incorrect values.

6.2.2 Harmonic oscillator

Data generation and inference model In this experiment we move towards a more realistic example involving the solution of the ODE that corresponds to a harmonic oscillator:

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = \frac{F(t)}{m} \tag{6.13}$$

where m is the mass being displaced, ζ is the damping ratio, ω_0 the proper angular frequency and $F(t)$ is the driving force applied to the system. In our case the driving force is sinusoidal, in particular:

$$F(t) = F_0 \sin(\omega)$$

where ω is the driving frequency of the force and F_0 its amplitude. We generate the data by numerically solving the ODE and fixing the values $\omega = 2, m = 1.3, \zeta = 0.28, \omega_0 = 2.78$ and sampling uniformly between 9 and 11 ($F_0 \sim \mathcal{U}[9, 11]$). However, a closed form solution of the above equation is also available for the steady-state regime,

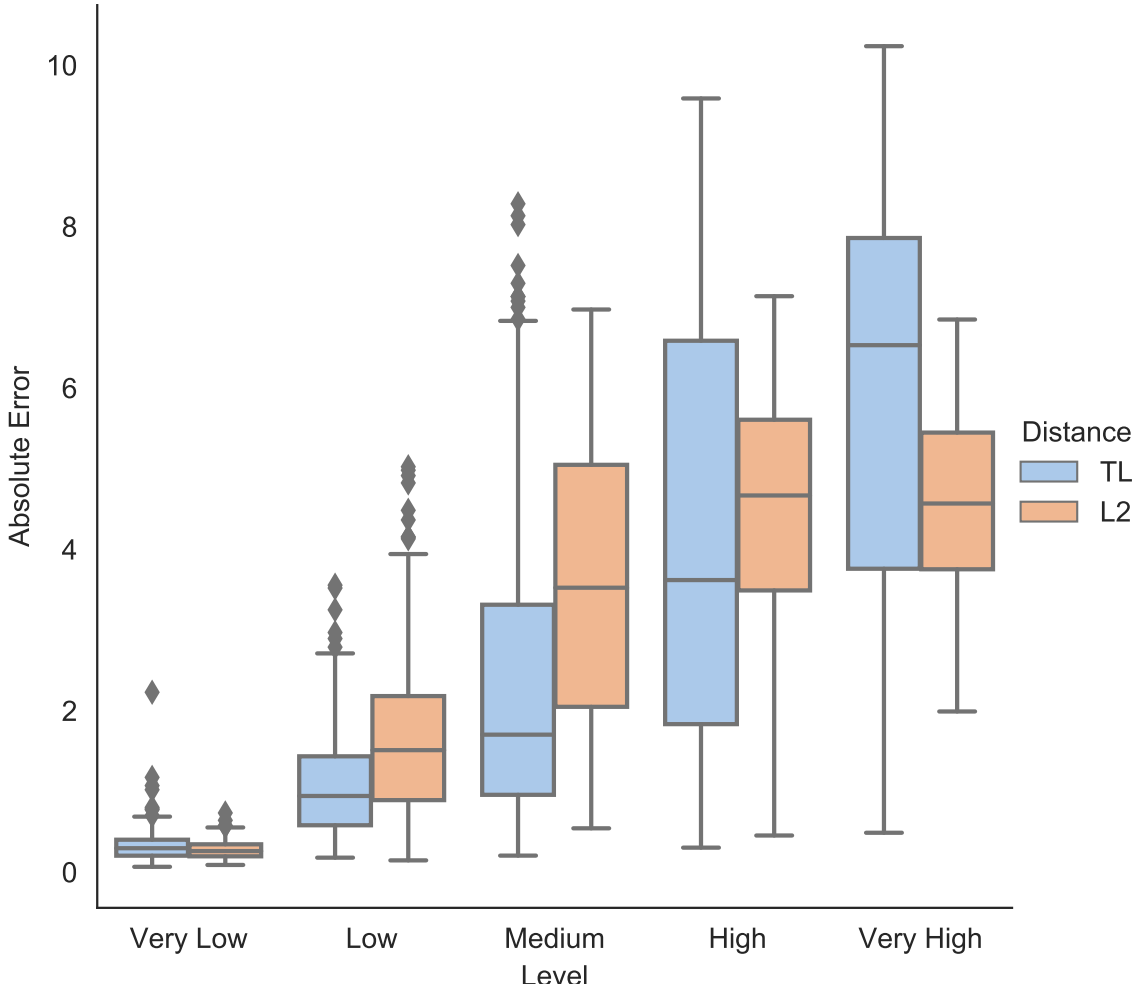


Figure 6-8: Box plot of absolute errors averaged across the model coefficients, for each intensity level

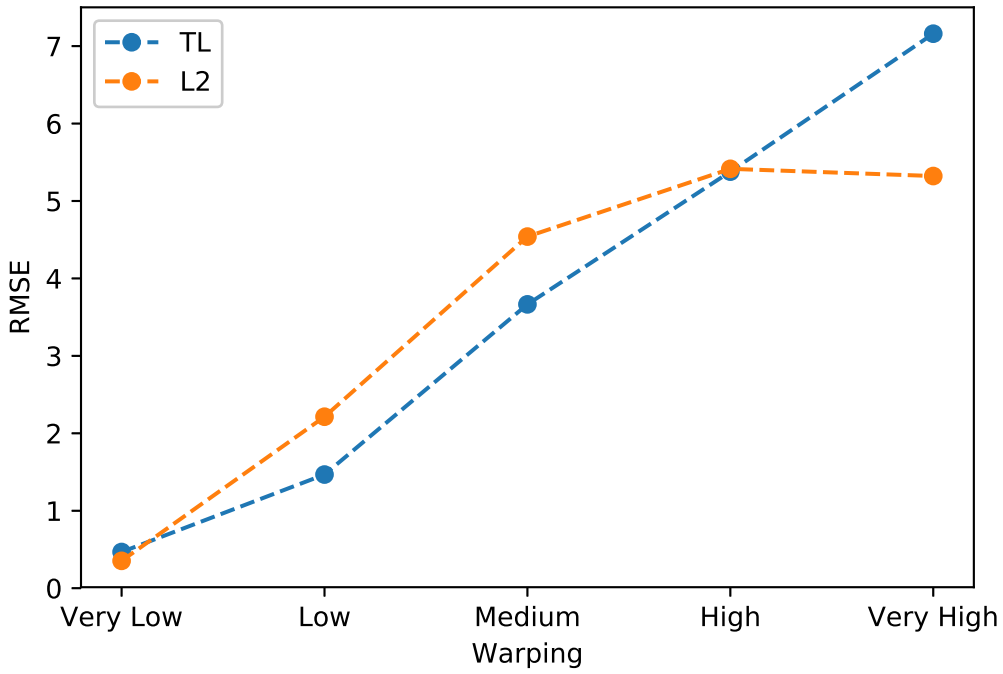


Figure 6-9: Average RMSE vs. warping intensity level for the TL and ℓ_2 linear regression

in particular:

$$x(t) = \frac{F_0}{mZ_m\omega}\sin(\omega t + \varphi) \tag{6.14}$$

where:

$$Z_m = \sqrt{(2\omega_0\zeta)^2 + \frac{1}{\omega^2}(\omega_0^2 - \omega^2)^2}$$
$$\varphi = \arctan\left(\frac{2\omega\omega_0\zeta}{\omega^2 - \omega_0^2}\right) + n\pi$$

where n is chosen such that the value of $\varphi \in [-\pi, 0]$. We therefore use this expression as a model for inversion, introducing some misspecification by neglecting to model the transitory regime. Through this model, we want to characterize the impedance Z_m of a system for different values of F_0 and fixed values of the other parameters , i.e., $\omega = 2$ and $m = 1.3$. We however assume $\varphi = 0$, even though this assumption does not hold for most systems, and its value is instead linked to Z_m. This simplifying assumption reduces the complexity of the inverse problem: by not having to estimate φ, we transition from a 2D non-linear problem, to a a 1D linear problem. Of course, the lack of appropriately modeled phase shift adds to the misspecification already introduced by neglecting the modeling of the transitory phase. Through this experiment we want to test whether the TL-based regression proves to be more robust in recovering the correct value of the impedance, which in our case is $Z_m = 2.40$. We report in Figure 6-10 the numerical solution with added noise, superposed with the theoretical steady-sate solution and the one resulting from the misspecified model, calibrated at the right value of Z_m. As it can be seen from the image, the misspecification at steady-state reduces to a simple phase shift. We acknowledge the fact, in this specific instance, the issue could be resolved by inferring φ at the same time as Z_m. However, this would mean making the problem non-linear and increase its dimension. For the harmonic oscillator this may not pose significant challenges, but in general enlarging the search space (especially in a non-linear model) could pose parameter identifiability issues or ill-posedness. We therefore use this simple physics-inspired system as a test bed for our methodology and motivation for more complex problems.

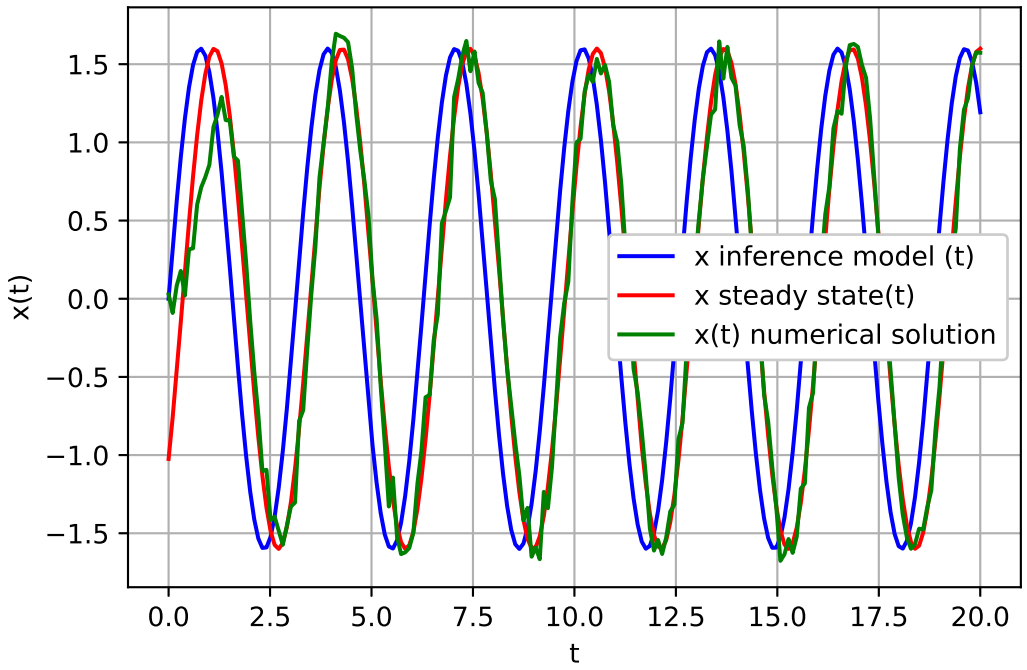


Figure 6-10: Average RMSE vs. warping intensity level for the TL and ℓ_2 linear regression

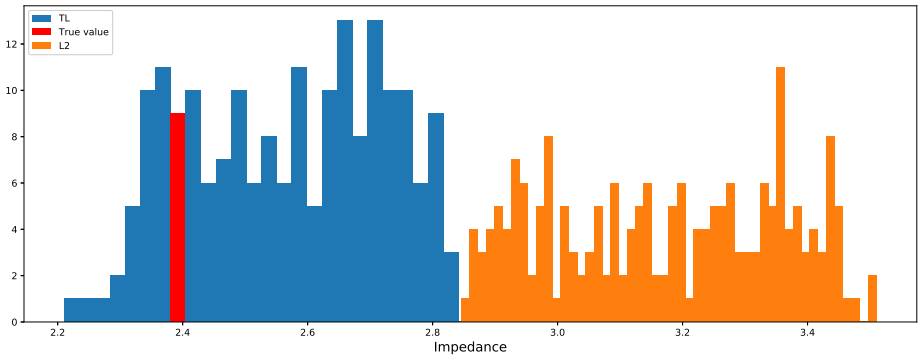


Figure 6-11: Histograms of recovered impedance through TL and ℓ_2-based linear regression, for different values of F_0. $\ell_{2_{RMSE}} = 0.79 > TL_{RMSE} = 0.26$.

Results We now report the histograms of the recovered values for the impedance Z_m for each of the $n_{samples} = 200$ of driving force amplitude F_0 in Figure 6-11. As it can be clearly seen, the TL is able to overcome the misspecification introduced on φ and produce values that are on average closer to the truth. In particular we have an RMSE associated with the ℓ_s of 0.79 vs. a TL one of 0.26.

6.2.3 Seismic wave

Data generation and inference model In this last experiment we test the TL-regression model against the problem of moment tensor inversion. While not as as extensive as the experiments performed in Chapter 4, we will use some of those waveforms to check whether the use of the TL-distance can be of benefit in the context of misspecified deterministic inversion. For our experiment we will generate data by randomly sampling some i.i.d Gaussian noise and add it to the overthrust model generated waveform for a specific station (NW) and displacement component (Z). As an inference model we will instead use the corresponding layered media model generated waveform as in section 4.2. For clarity the data will be generated as:

$$\mathbf{y} \sim \mathbf{G}^{Z}_{NW}(\mathbf{x}_{\text{true}}, \mathbf{V}_{3\text{D}}, \mathbf{t}) \cdot \mathbf{m}^{T}_{s_{\text{true}}} + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \tag{6.15}$$

while the model used for inference will be:

$$\mathbf{u} = \mathbf{G}^{Z}_{NW}(\mathbf{x}_{\text{true}}, \mathbf{V}_{\text{2D}}, \mathbf{t}) \cdot \mathbf{m}^{T}_{s_{\text{true}}} \tag{6.16}$$

where $\mathbf{x}_{\text{true}}$ and $\mathbf{m}_{s_{\text{true}}}$ are identical to those used in section 4.2. We observed in this instance that the addition of a regularization term is necessary to control the magnitude of the recovered set of moment tensor and reproduce, at least partially, the role played by the bounded uniform prior used in the Bayesian setting. In particular, opting for ridge regression (with parameter γ), the analytical solution for a least squares problem [63] (fixing the permutation $\boldsymbol{\sigma}$ can be written as:

$$\hat{\boldsymbol{m}}_{\boldsymbol{\sigma}} = \left(\boldsymbol{\Phi}(\mathbf{x})^{T}\boldsymbol{\Phi}(\mathbf{x}) - \gamma\mathbb{I}\right)^{-1}\boldsymbol{\Phi}(\mathbf{x})^{T}\mathbf{y}_{\boldsymbol{\sigma}} \tag{6.17}$$

By substituting this expression in equation (6.7) and redefining the matrix $\mathbf{S}$ accordingly:

$$\mathbf{A} = \boldsymbol{\Phi}(\mathbf{x})\left(\boldsymbol{\Phi}(\mathbf{x})^{T}\boldsymbol{\Phi}(\mathbf{x}) - \gamma\mathbb{I}\right)^{-1}\boldsymbol{\Phi}(\mathbf{x})^{T}$$

we revert to equation 6.9. The rest of the algorithm and computational solution is not impacted.

Results We first report the results for the γ that achieves the lowest RMSE: $\gamma = 1$. In Figure 6-12 we can observe how the TL is generally close to the true values except for one moment tensor component. The plots for the remaining values of γ can be found in Figures 6-13, 6-14, 6-15, 6-16, 6-17.

We now report in table 6.2 below the differences between ℓ_2 and TL-based RMSE, as a function of γ. If the TL exhibits a lower RMSE we highlight the difference (positive) in green; otherwise (negative) in red. We remind readers that although the highest discrepancy is observed for $\gamma = 0.8$, the associated RMSE was lower both for the TL and ℓ_2 for $\gamma = 1$.

Ridge	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	$\Delta^{RMSE}_{\ell_2-\text{TL}}$	RMSE
Param.	m_{ee}	m_{ne}	m_{ez}	m_{nn}	m_{nz}	m_{zz}	Average	TL
$\gamma = 0$	329	395	-47	323	-139	2.08	143	488
$\gamma = 0.5$	2.59	0.41	0.07	2.17	0.02	2.88	1.35	0.49
$\gamma = 0.8$	0.25	0.44	0.13	-0.28	0.08	0.15	0.13	0.34
$\gamma = 0.9$	0.12	0.39	0.12	-0.43	0.066	0.023	0.05	0.34
$\gamma = 1$	0.11	0.34	0.13	-0.50	0.07	0.02	0.03	0.32
$\gamma = 2$	0.19	-0.06	0.02	-0.23	0.02	-0.07	-0.02	0.38
$\gamma = 10$	-0.26	-0.04	0.54	-0.16	0.19	-0.42	-0.03	0.47

Table 6.2: RMSE scores and relative differences for different values of Ridge parameter between TL and ℓ_2 based regression.

While the regularization contributed to the improvement of the results, the solution appears fairly sensitive to it.

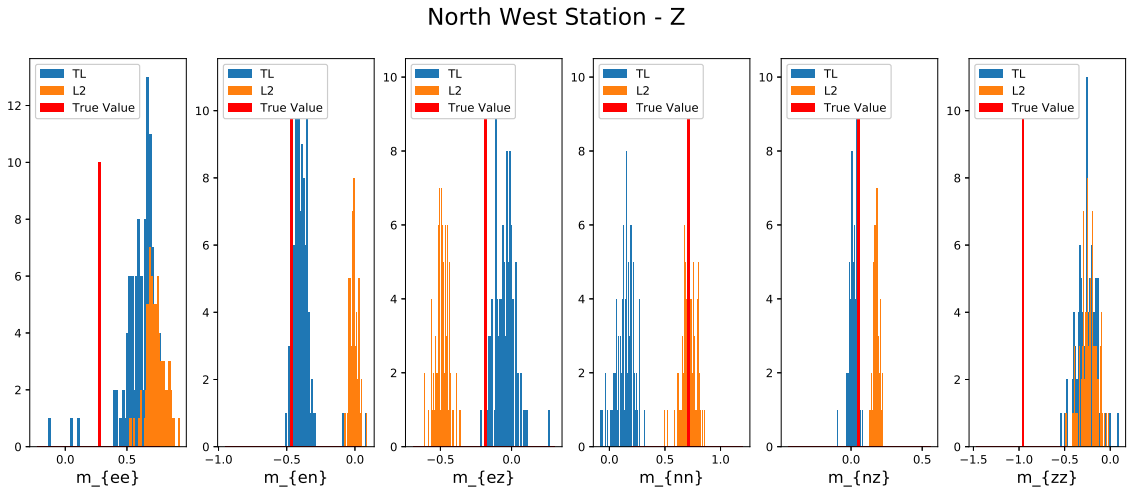


Figure 6-12: Histogram of recovered moment tensor for $n_{samples} = 1000$ of Gaussian noise $\gamma = 1$

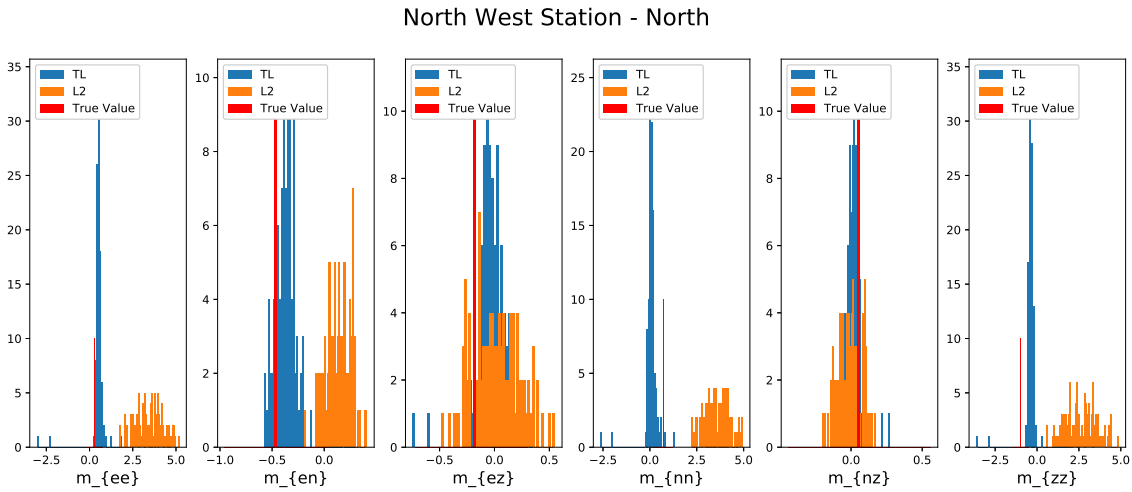


Figure 6-13: Histogram of recovered moment tensor for $n_{samples} = 1000$ of Gaussian noise $\gamma = 0.5$

Remarks on comparison with Bayesian inversion

We would like to close this section with a few remarks on the comparison between the Bayesian inversion results obtained in Chapter 4 and the deterministic inversion results just described. Generally speaking, it appears that advantage brought by the TL is less prominent than the one shown in Chapter 4. However, rather than taking these experiments as a final assessment on the performance of the TL in a deterministic regression setting, we would like to list a few factors that, while not fully considered in our study, may significantly enhance its performance. In particular:

1. a more extensive use of the regularization term and fine-tuning of relative parameters may lead to better results. In Bayesian inversion, this role was played by the bounded uniform prior, which is not easily translatable in the regression formulation without adding further constraints;

2. while in the Bayesian formulation the inversion could rely on data coming from 7 stations with a waveform for each sense of displacement (21 waveforms in total), in this regression we considered one waveform only. This choice was made to speed up computation since adding more waveforms would have required solving for 20 additional permutation matrices. The formulation presented is however

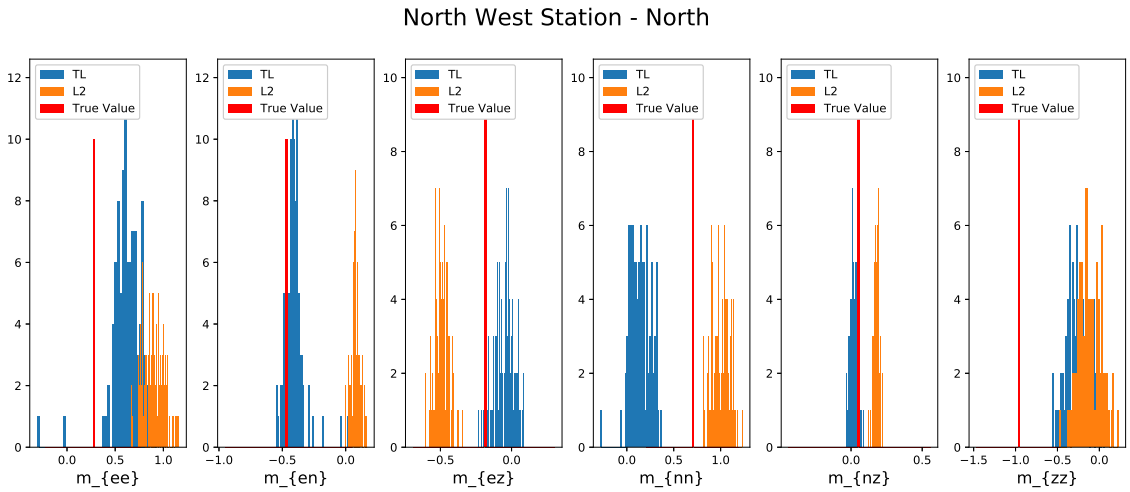


Figure 6-14: Histogram of recovered moment tensor for $n_{samples} = 1000$ of Gaussian noise $\gamma = 0.8$

still valid and more data could allow results to be better constrained;

3. the role played by the prior on the s parameter in the Gibbs posterior also has an impact in the Bayesian setting, which is difficult to relate to anything specific to the regression formulation.

6.3 Conclusion

In this chapter we presented an alternative formulation for linear regression with permutations of the data vector. We described in what way our formulation differs from previously proposed ones together with a dedicated algorithm to solve it. We tested the advantages of such a formulation in a misspecified context on a number of benchmark problems. More investigation is needed for faster and exact algorithms for the solution of the problem, as well as additional regularization terms for the objective function.

Acknowledgements The work presented in this chapter was performed in collaboration with Jean Pauphilet (London Business School.).

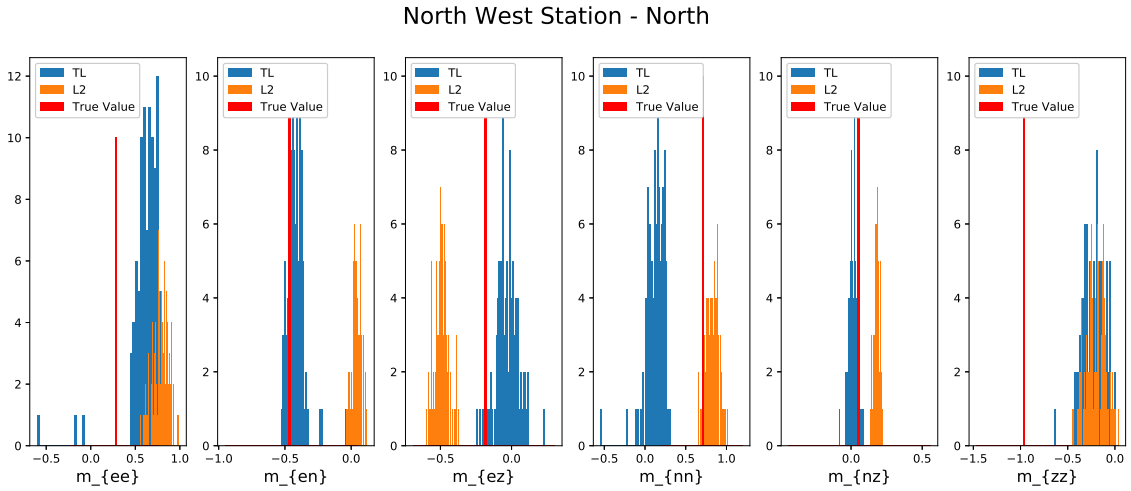


Figure 6-15: Histogram of recovered moment tensor for $n_{samples} = 1000$ of Gaussian noise $\gamma = 0.9$

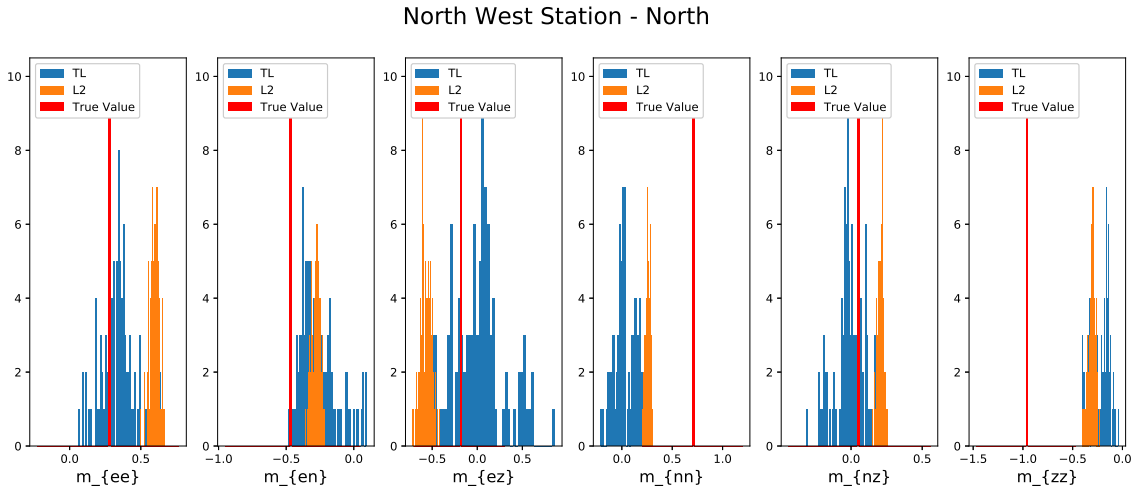


Figure 6-16: Histogram of recovered moment tensor for $n_{samples} = 1000$ of Gaussian noise $\gamma = 2$

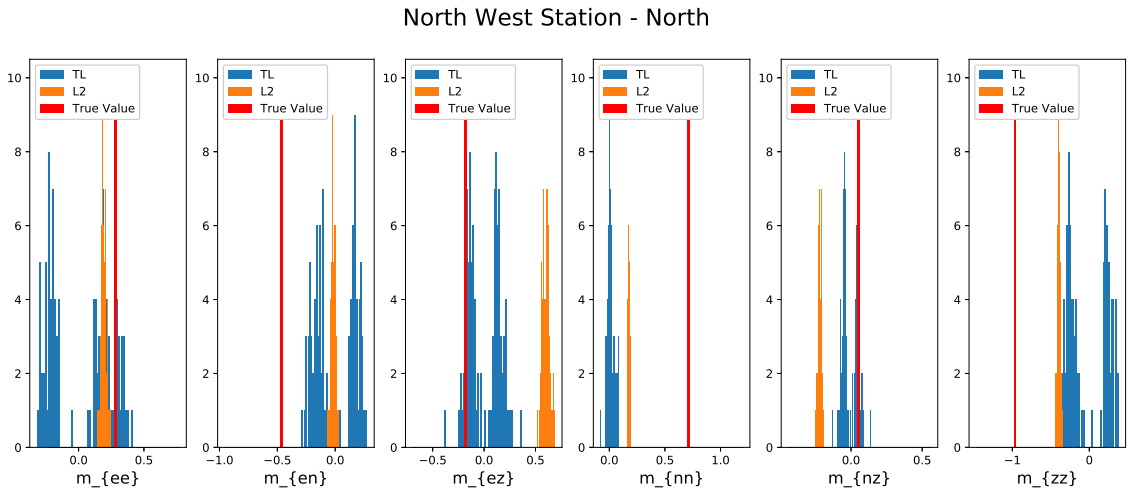


Figure 6-17: Histogram of recovered moment tensor for $n_{samples} = 1000$ of Gaussian noise $\gamma = 10$

Chapter 7

Conclusions and future directions

The issue of model misspecification is a longstanding one in inverse problems and modeling in general. In this thesis, we moved away from classical approaches to this problem, which included either enriching the statistical or physical complexity of the model or some "loss" minimizing approach (e.g., safe Bayes). We instead identified a pathway towards reducing the impact of model misspecification, i.e., achieving *robustness*, through the use of alternative misfit measures. In particular, misfit functions tailored to the the specific data-model pair being considered and possible ways it may be affected by model error. We restricted our attention to the use-case of moment tensor inversion in seismic inverse problems. For this setting and specific type of data (i.e., time series), we identified the transport-Lagrangian distance (an optimal transport-based distance) as valid misfit measure to reduce the impact of errors in the velocity models for seismic waves propagation. Aside from the specific application (SEG-EAGE Overthrust models) (Chapter 4), the investigation of the statistical properties of the TL distance (Chapter 5) and its application in a deterministic linear regression context (Chapter 6). We believe the research and results described in this thesis can be summarized as:

- showing the benefit of using optimal transport based distances in misspecified inverse problems involving time-series data;

- introducing a consistent framework for the adoption of such distances in a Bayesian context;
- a realistic application of the methodology in the field of moment tensor estimation in seismic inverse problems;
- a discussion of the statistical properties of the TL-distance;
- a new formulation of the liner regression problem with permutations of the data vector.

Some limitations of the work conducted in this thesis are:

1. the applicability of optimal transport based misfit measures to a specific category of inverse problems involving data that is discretized in time and space and that can benefit from the flexibility offered by horizontal transport for amplitude matching;
2. the lack of a definitive answer to the problem of identifying a tractable expression for a TL-based likelihood function;
3. an extended validation of the proposed method to a variety of applications beyond seismic inversion.

Based on these considerations, we also envision potential extensions of the proposed framework, which we outline as potential future research directions in this chapter. The numerical and theoretical results obtained in this thesis point at two main possible threads worth of further investigation:

1. Misfit measures for functional data;
2. Systematic identification and construction of misspecification-robust misfit measures - data-feature based projection operators;

In the next sections we will provide more details on each of the first two threads.

7.1 Misfit measures for functional data

Part of the motivation for looking into optimal transport as an alternative misfit measure for seismic inverse problems comes from the fact there are currently no distances that are able to capture the differences or similarities in *shape* between two signals, functions or geometric entities. While we already extensively described this problem in section 2, we want to propose here an extension of the transport-Lagrangian distance that builds on the already well-established construct of time-embedding (or delay reconstruction) with Wasserstein distances [12]. In particular, if the classic TL-distance between two vectors of equal length is defined as:

$$\mathrm{TL}_\lambda^P = \min_{\boldsymbol{\sigma}\in\mathrm{Perm}(n)} \sum_{i=1}^{n} |x_i - x_{\sigma(i)}|^p + \lambda |f(x_i) - g(x_{\sigma(i)})|^p = \tag{7.1}$$

$$= \min_{\boldsymbol{\sigma}\in\mathrm{Perm}(n)} ||\mathbf{x} - \mathbf{x}_\sigma||_p^p + ||f(\mathbf{x}) - g(\mathbf{x}_\sigma)||_p^p \tag{7.2}$$

we propose that instead of taking the squared difference between the indices x_i and $x_{\sigma(i)}$ and the value that the signals take at those indices only $|f(x_i) - g(an\sigma(i))|^2$, we extend this second contribution to be that of a window of length τ along the signals f and g being considered. By defining:

$$\tilde{f}_i = (f_i, f_{i-1}, f_{i-2}, \ldots, f_{i-\tau})_i \tag{7.3}$$

$$\tilde{g}_{\sigma(i)} = (f_{\sigma(i)}, f_{\sigma(i)-1}, f_{\sigma(i)-2}, \ldots, f_{\sigma(i)-\tau})_i \tag{7.4}$$

the time-embedded TL-distance then becomes:

$$\mathrm{TL}_\lambda^P = \min_{\boldsymbol{\sigma}\in\mathrm{Perm}(n)} \sum_{i=1}^{n} |x_i - x_{\sigma(i)}|^p + \lambda |\tilde{f}(x_i) - \tilde{g}(x_{\sigma(i)})|^p = \tag{7.5}$$

$$= \min_{\boldsymbol{\sigma}\in\mathrm{Perm}(n)} ||\mathbf{x} - \mathbf{x}_\sigma||_p^p + ||\tilde{f}(\mathbf{x}) - \tilde{g}(\mathbf{x}_\sigma)||_p^p \tag{7.6}$$

This delay reconstruction or time embedding formulation is supposed to better capture the concept of shape or curvature that is intrinsic to the ideal of functional data or signal. while in this thesis we did not test this newly defined distance extensively,

we would like to underscore that from a computational point of view this does not represents an increased cost since the only difference consists in the re-definition of the cost associated to a specific permutation.

7.2 Data-feature based projection operators

Although in this thesis we focused on transport-Lagrangian distances and the role that optimal transport can play in reducing the impact of model misspecification for time series data, it is possible to envision a broader perspective in which the concept of *transforming, filtering or projecting* the data-model misfit is systematically exploited.

To make some of these ideas more concrete, let us suppose we are interested in a phenomenon that is described through a deterministic model u that is a function of some parameter vector $\theta \in \Theta$ such that $\mathbf{u}_{model} = \mathbf{u}(\theta)$. Let us also suppose that we are in a full Bayesian setting and thus the parameters are endowed with a prior distribution $p(\theta)$ and the observations are related to the model parameters through a likelihood function $f(\mathbf{y}|\theta)$. We assume that the observations $\mathbf{y}$ are generated by a distribution $g(\mathbf{y})$ that is different from f and we are in presence of model misspecification. We make the additional assumption that the misfit between g and f is linked to the poor modeling of a certain aspect of the total phenomenon and that only a subset of the parameters is associated with it. We therefore divide the parameter set into two subsets: parameters not associated to model misspecification (WS - "well specified") and parameters associated with model misspecification (MS - " well specified"):

$$\Theta = \{\theta_{WS}, \theta_{MS}\}; \tag{7.7}$$

We want to look for a data transformation $\mathcal{L}(\mathbf{y})$ that achieves the following objectives:

$$\exists \theta^* \quad \text{s.t.} \quad \mathcal{L}(\mathbf{y}) \sim f(\cdot|\theta^*), \tag{7.8}$$

$$\mathcal{I}(\mathcal{L}(\mathbf{y}), \theta_{MS}) \approx 0, \tag{7.9}$$

$$\mathcal{I}(\mathcal{L}(\mathbf{y}), \theta_{WS}) \approx \mathcal{I}(\mathbf{y}, \theta_{WS}). \tag{7.10}$$

Let us discuss each of the criteria above:

1. The first objective (7.8) identifies the "projection" condition, i.e., the fact that the transformation shall bring the observed data to the sample space described by the model distribution f. Indeed, if we assume that the model is correctly specified with respect to θ_{WS}, then, after the transformation there shall always be a θ such that the distribution of the transformed data and the parameterized one are equal. In this context the Berstein-Von-Mises theorem also ensures us that the inference we will make of θ_{WS} will be asymptotically correct;

2. The transformation $\mathcal{L}(\mathbf{y})$ is a transformation that depends on the model being used for inference, thus $\mathcal{L}(\mathbf{y}) = \mathcal{L}_f(\mathbf{y})$;

3. The mutual information ($\mathcal{I}$) criteria aim at addressing the "quality" of the transformation we are seeking. While it is always possible to find some sort of transformation such that the data belongs to the model space, the challenge is to achieve this objective without loosing or distorting the information needed in order to infer θ_{WS} correctly. These criteria can be used to assess: 1) whether the chosen transformation is decreasing the amount of information related to θ_{WS}; 2) how much information is still carried relative to θ_{MS} (ideally minimal, since θ_{MS} is now treated as a nuisance parameter).

4. The criteria above are those a *perfect* transformation should satisfy. In practice, it is not straightforward to identify the transformation that meets all of the above requirements fully. In addition we must also remember that no model is ever fully well-specified, and that therefore even for θ_{WS} we will not reach condition

7.8 perfectly. All the above criteria are therefore to be intended "loosely" as a way to test a transformation that is built off some specific knowledge about the problem.

We conclude this section by listing some ideas on how to identify these transformations for the test case of seismic inversion.

7.2.1 A test-case: moment tensor inversion

We have a model $\mathbf{u}$ that is used to predict observed waveforms $\mathbf{y}$ in the following fashion:

$$\mathbf{y} = \mathbf{u}(\mathbf{x}_{\text{true}}, \mathbf{V}, \mathbf{m}_{\text{true}}, \mathbf{t}) + \mathbf{e} \quad \text{where: } \mathbf{e} \sim \mathcal{N}(0, \sigma^2) \tag{7.11}$$

$$\mathbf{u} = \mathbf{G}(\mathbf{x}_{\text{true}}, \mathbf{V}, \mathbf{t}) \cdot \mathbf{m}^T \tag{7.12}$$

We know that $\mathbf{V}$ is highly misspecified and therefore affects the quality of the inversion on $\mathbf{m}$ as well.

We claim that there exist some data features related to $\mathbf{m}$ that are independent of the specific $\mathbf{V}$ that has generated the data and that can be used to perform more robust inference of $\mathbf{m}$. In this thesis we have proven the benefits of using the TL-distance as a misfit measure for problems like the one we just described. Our likelihood function will therefore be expressed through a the TL-distance. As a reminder:

$$p(\text{TL}(\mathbf{y}, \mathbf{u})^2 | \mathbf{m}, \mathbf{V}) \tag{7.13}$$

Instead of transforming the data directly, we instead propose to act on the mappings $\mathcal{C}$ (i.e., permutation matrices) that we obtain we calculating the TL-distance between two waveforms. In other words, we "register" all the mapping $\mathcal{C}$ between the misspecified data and the model to those that we would obtain if the data were coming from a model set at $\mathbf{V_o}$. This allows to "wash the data out" of any features relevant to the

misspecified $\mathbf{V}$ and infer $\mathbf{m}$ in a well specified setting.

The foreseen methodology would be articulated in three steps:

1. **Explore** the model both in a well specified and misspecified context;
2. **Build** the mapping transformation or mapping between misspecified mappings and well specified ones;
3. **Transform** the misspecified mappings into well-specified ones while performing inference with a given data-set.

We now provide some ideas and possible strategies to follow for each of these steps.

Explore Sample M pairs of $(\mathbf{m}_i, \mathbf{m}_j)$ from an appropriate prior and calculate the respective waveforms:

$$\mathbf{u}_i(\mathbf{t}) = \mathbf{u}(\mathbf{V} = \mathbf{V_o}, \mathbf{m} = \mathbf{m}_i, \mathbf{t}) \tag{7.14}$$

$$\mathbf{u}_j(\mathbf{t}) = \mathbf{u}(\mathbf{V} = \mathbf{V_o}, \mathbf{m} = \mathbf{m}_j, \mathbf{t}) \tag{7.15}$$

where $\mathbf{u} \in \mathbb{R}^N$. For each pair i, j, calculate:

$$[\mathcal{D}_{i,j}, \mathcal{C}_{i,j}] = TL_2(\mathbf{u}_i, \mathbf{u}_j) \text{ for: } i, j = 1 \dots N; \tag{7.16}$$

Where:

$$\begin{aligned} \mathcal{D}_{i,j} &= \text{ total distance between waveform } \mathbf{u}_i = \mathbf{u}((\mathbf{m}, \mathbf{V})_i) \text{ and } \mathbf{u}_j = \mathbf{u}((\mathbf{m}, \mathbf{V})_j) \\ \mathcal{C}_{i,j} &= \text{ transport maps between waveform } \mathbf{u}_i = \mathbf{u}((\mathbf{m}, \mathbf{V})_i) \text{ and } \mathbf{u}_j = \mathbf{u}(\mathbf{m}, \mathbf{V})_j) \end{aligned}$$

With this notation we simply want to emphasize that not only the numerical value of the distance will be considered, but also the mapping (i.e., optimal permutation matrix) that led to it. Since the cost maps $\mathcal{C}$ calculated in this way come from a well

specified model, we will refer to them as:

$$\mathcal{C}^{\mathrm{V}}_{i,j} \tag{7.17}$$

and we will group them in a set $\boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}}$. Then, following the same logic, we calculate the maps between the modeled waveforms and the actual data $\mathbf{y}$, which could come, for example, from a model simply set at a $\mathbf{V} \neq \mathbf{V}_o$. The maps build this way would be referred to as misspecified $\boldsymbol{\mathcal{C}}^{\not{\mathrm{V}}}$.

Build At this stage it is necessary to identify the transformation $\mathcal{L}(\cdot)$ that would allow the kind of data-projection sought at the beginning of this chapter. In general terms, by borrowing some language from linear algebra, we could describe this transformation as follows:

- identify the span of the $\boldsymbol{\mathcal{C}}^{\mathrm{V}}$;
- identify the span of the $\boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}}$;
- identify a projector $\mathcal{L}(\cdot) = \Gamma = \langle \boldsymbol{\mathcal{C}}^{\mathrm{V}}, \boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}} \rangle$.

Of course the notion of span$\{\boldsymbol{\mathcal{C}}^{\mathrm{V}}\}$ or span$\{\boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}}\}$ are not immediately defined. A path that we started to explore required re-defining $\boldsymbol{\mathcal{C}}^{\mathrm{V}}$ and $\boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}}$ as matrices (which can be done by expressing permutations as vectors) and then seeking a linear map between the two spaces via positive least squares, for e.g.:

$$\min_{\Gamma \in \mathbb{R}^{M \times N}_{\geq > 0}} ||\boldsymbol{\mathcal{C}}^{\mathrm{V}} - \boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}} \Gamma||_F \tag{7.18}$$

or common-subspace method, by first calculating the singular value decomposition of:

$$\boldsymbol{\mathcal{C}}^{\mathrm{V}} = U^{\mathrm{V}} C V^T \tag{7.19}$$

$$\boldsymbol{\mathcal{C}}^{V_{\mathrm{misp}}} = U^{\not{\mathrm{V}}} S V^T \tag{7.20}$$

and then expressing Γ as:

$$\Gamma = \frac{VCSV^T}{||VCSV^T||_F} \tag{7.21}$$

We briefly experimented these options and obtained results that required further refining or pointed towards the construction of non-linear operators (e.g., through neural networks).

Appendix A

Conditions to apply the Lyapunov central limit theorem

We want to show that the Lyapunov central limit theorem can be applied to the sum of the set of non-central chi-squared random variables, similarly to those defined in section 5.1.4. To simplify the notation, we define a Z_k as the sum of n non-central chi-squared random variables with one degree of freedom:

$$Z_k = \sum_{i=1}^{n} \chi_i^2 \tag{A.1}$$

We assume these n variables are independent (consistently with the setup in section 5.1.4)), but have different variance and mean. In fact:

$$\mathbb{E}(\chi_i^2) = 1 + \mu_i^2 \text{ and } \mathbb{V}\text{ar}(\chi_i^2) = \sigma_i^2 = 2(1 + 2\mu_i^2) \tag{A.2}$$

In order to apply Lyapunov CLT we must show that that random variables $|\chi_l^2|$ have moments of some order $2+\delta$ and that the rate of growth of these moments is bounded in terms of the Lyapunov condition (theorem 27.3 in [14]):

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\Big[\big|\chi_i^2 - \mathbb{E}[\chi_i^2]\big|^{2+\delta}\Big] = 0 \tag{A.3}$$

where, in our case $s_n^2 = 2n + 4\sum_{i=1}^{n} \mu_i^2$. By choosing $\delta = 2$:

$$\lim_{n\to\infty} \frac{1}{s_n^4} \sum_{i=1}^{n} \mathbb{E}\Big[\big|\chi_i^2 - \mathbb{E}[\chi_i^2]\big|^4\Big] \tag{A.4}$$

$$= \lim_{n\to\infty} \frac{1}{s_n^4} \sum_{i=1}^{n} \sigma_i^4 \mathbb{E}\left[\left|\frac{\chi_i^2 - \mathbb{E}[\chi_i^2]}{\sigma_i}\right|^4\right] \tag{A.5}$$

we recognize in the external expected value the kurtosis (i.e. standardized fourth-moment) for a non-central chi-squared random variable, which gives [117]:

$$\mathbb{E}\left[\left|\frac{\chi_i^2 - \mathbb{E}[\chi_i^2]}{\sigma_i}\right|^4\right] = \frac{12(1 + 4\mu_i^2)}{(1 + 2\mu_i^2)^2} \tag{A.6}$$

By substituting this expression and the value of s_n in (A.5), we obtain:

$$= \lim_{n\to\infty} \frac{1}{s_n^4} \sum_{i=1}^{n} \sigma_i^4 \mathbb{E}\left[\left|\frac{\chi_i^2 - \mathbb{E}[\chi_i^2]}{\sigma_i}\right|^4\right] \tag{A.7}$$

$$= \lim_{n\to\infty} \frac{1}{\left(2n + 4\sum_{i=1}^{n} \mu_i^2\right)^2} \sum_{i=1}^{n} 4(1 + 2\mu_i^2)^2 \frac{12(1 + 4\mu_i^2)}{(1 + 2\mu_i^2)^2} \tag{A.8}$$

$$= \lim_{n\to\infty} \frac{48n + 192\sum_{i+1}^{n} \mu_i^2}{\left(2n + 4\sum_{i=1}^{n} \mu_i^2\right)^2} = 0 \tag{A.9}$$

Appendix B

Covariance matrix for X_l

The covariance matrix for the chi-squared non-central random variables is structured as follows:

$$\mathbb{C}\text{ov}(X_{l'}, X_{l''}) = \begin{cases} \mathbb{V}\text{ar}(X_l) = 2 + 4\mu_l^2 & \iff l' = l'' \\ 0 & \iff l' \neq l'' \text{ and } i_{l'} \neq i_{l''} \\ \beta & \iff l' \neq l'' \text{ and } i_{l'} = i_{l''} \end{cases} \tag{B.1}$$

The exact expression of β is derived in this appendix. Let:

$$X_{ij} = \left(\frac{g(t_j) - f(t_i)}{\sigma}\right)^2 = \left(\frac{g(t_j) - \mu_{f_i}}{\sigma} - \zeta\right)^2 \tag{B.2}$$

$$X_{pq} = \left(\frac{g(t_q) - f(t_p)}{\sigma}\right)^2 = \left(\frac{g(t_q) - \mu_{f_p}}{\sigma} - \zeta\right)^2 \tag{B.3}$$

Where $\zeta \sim \mathcal{N}(0,1)$. Calling $\frac{g(t_j)-\mu_{f_i}}{\sigma} = a_{ij}$ and $\frac{g(t_q)-\mu_{f_p}}{\sigma} = a_{pq}$, we have:

$$X_{ij} = a_{ij}^2 + \zeta^2 - 2a_{ij}\zeta \tag{B.4}$$

$$X_{pq} = a_{pq}^2 + \zeta^2 - 2a_{pq}\zeta \tag{B.5}$$

consequently:

$$X_{ij}X_{pq} = a_{ij}^2 a_{pq}^2 + a_{ij}^2\zeta^2 - 2a_{ij}^2 a_{pq}\zeta \tag{B.6}$$

$$+ a_{kl}^2\zeta^2 + \zeta^4 - 2a_{kl}\zeta^3 \tag{B.7}$$

$$- 2a_{ij}a_{kl}^2\zeta - 2a_{ij}\zeta^3 + 4a_{ij}a_{pq}\zeta^2 \tag{B.8}$$

Recalling:

$$\mathbb{E}(\zeta) = 0 \tag{B.9}$$

$$\mathbb{E}(\zeta^2) = 1 \tag{B.10}$$

$$\mathbb{E}(\zeta^3) = 0 \tag{B.11}$$

$$\mathbb{E}(\zeta^4) = 3 \tag{B.12}$$

We have:

$$\mathbb{E}[X_{ij}] = a_{ij}^2 + 1 \tag{B.13}$$

$$\mathbb{E}[X_{pq}] = a_{pq}^2 + 1 \tag{B.14}$$

$$\mathbb{E}[X_{ij}X_{pq}] = a_{ij}^2 a_{pq}^2 + a_{ij}^2 + a_{pq}^2 + 3 + 4a_{ij}a_{pq} \tag{B.15}$$

Therefore:

$$\mathbb{C}\text{ov}(X_{ij}, X_{pq}) = \mathbb{E}[X_{ij}X_{pq}] - \mathbb{E}[X_{ij}]\mathbb{E}[X_{pq}] = 4a_{ij}a_{pq} + 2$$

Substituting the values of a_{ij} and a_{pq}:

$$\beta = \mathbb{C}\text{ov}(X_{ij}, X_{pq}) = \frac{4}{\sigma^2}(g(t_j) - \mu_{f_i})(g(t_q) - \mu_{f_p}) + 2 \tag{B.16}$$

As a double check, we can verify that if $i = p$ and $j = q$, we have:

$$\mathbb{C}\text{ov}(X_{ij}, X_{ij}) = \frac{4}{\sigma^2}(g(t_j) - \mu_{f_i})^2 + 2 = 4\mu_l^2 + 2 = \mathbb{V}\text{ar}(X_{ij}) \tag{B.17}$$

Bibliography

[1] A. Abid, A. Poon, and J. Zou. Linear regression with shuffled labels. *arXiv preprint arXiv:1705.01342*, 2017.

[2] K. Aki. Evidence for magma intrusion during the mammoth lakes earthquakes of may 1980 and implications of the absence of volcanic (harmonic) tremor. *Journal of Geophysical Research: Solid Earth*, 89(B9):7689–7696, 1984.

[3] K. Aki and P. G. Richards. *Quantitative seismology*. 2002.

[4] P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

[5] F. Aminzadeh, N. Burkhard, T. Kunz, L. Nicoletis, and F. Rocca. 3-d modeling project: 3rd report. *The Leading Edge*, 14(2):125–128, 1995.

[6] F. Aminzadeh, N. Burkhard, J. Long, T. Kunz, and P. Duclos. Three dimensional seg/eaeg models—an update. *The Leading Edge*, 15(2):131–134, 1996.

[7] F. Aminzadeh, N. Burkhard, L. Nicoletis, F. Rocca, and K. Wyatt. Seg/eaeg 3-d modeling project: 2nd update. *The Leading Edge*, 13(9):949–952, 1994.

[8] A. Amir, B. Romain, G. Stephane, A. Francois, T. Pierre, and V. Jean. Regularized seismic full waveform inversion with prior model information. *Geophysics*, 78(2):R25–r36, March-April 2013. doi: 10.1190/GEO2012-0104.1.

[9] G. Antoine, A. Gboyega, and D. Esteban. Constrained full-waveform inversion by model reparameterization. *Geophysics*, 77(2):R117–r127, March-April 2012. doi: 10.1190/GEO2011-0196.1.

[10] B. Arild and O. Henning. Bayesian linearized avo inversion. *Geophysics*, 68(1):185–198, January-February 2003. doi: 10.1190/1.1543206.

[11] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal processing*, 18(4):349–369, 1989.

[12] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Inference in generative models using the wasserstein distance. *arXiv preprint arXiv:1701.05146*, 1(8):9, 2017.

[13] D. P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

[14] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[15] C. M. Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

[16] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[17] M. Bouchon and K. Aki. Discrete wave-number representation of seismic-source wave fields. *Bulletin of the Seismological Society of America*, 67(2):259–277, 1977.

[18] E. Bozdağ, J. Trampert, and J. Tromp. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870, 05 2011.

[19] R. Brossier, S. Operto, and J. Virieux. Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3):R37–r46, 2010.

[20] R. Brossier, S. Operto, and J. Virieux. Velocity model building from seismic reflection data by full-waveform inversion. *Geophysical Prospecting*, 63(2):354–367, 2015.

[21] S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: an integral part of inference. *Biometrics*, pages 603–618, 1997.

[22] B. Carey, S. Fatimetou, Z. S., and C. G. Multiscale seismic waveform inversion. *Geophysics*, 60(5):457–1473, September-October 1995. doi: 10.1093/gji/ggt258.

[23] A. D. Chan, M. M. Hamdy, A. Badre, and V. Badee. Wavelet distance measure for person identification using electrocardiograms. *IEEE transactions on instrumentation and measurement*, 57(2):248–253, 2008. doi: 10.1109/TIM.2007.909996.

[24] C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 158(3):419–444, 1995.

[25] F. Chen and D. Peter. Constructing misfit function for full waveform inversion based on sliced wasserstein distance. In Ifema, editor, *80th EAGE Conference and Exhibition 2018*, Madrid, Spain, June 2018. Ifema. doi: 10.3997/2214-4609.201801030.

[26] P. F. Chen, M. Nettles, E. A. Okal, and G. Ekström. Centroid moment tensor solutions for intermediate-depth earthquakes of the wwssn–hglp era (1962–1975). *Physics of the Earth and Planetary Interiors*, 124(1-2):1–7, 2001.

[27] S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

[28] O. Coutant. Programme de simulation numerique axitra, rapport lgit. *Univ. Joseph Fourier, Grenoble*, 1990.

[29] E. Crase, A. Picat, M. Noble, J. McDonald, and A. Tarantola. Robust elastic nonlinear waveform inversion: Application to real data. *Geophysics*, 55(5):527–538, May 1990.

[30] H. A. David and H. N. Nagaraja. *Order statistics*. Wiley Online Library, 2004.

[31] L. P. De Figueiredo, D. Grana, M. Santos, W. Figueiredo, M. Roisenberg, and G. S. Neto. Bayesian seismic inversion based on rock-physics prior modeling for the joint estimation of acoustic impedance, porosity and lithofacies. *Journal of Computational Physics*, 336:128–142, 2017. doi: 10.1016/j.jcp.2017.02.013.

[32] P. Diaconis and D. Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

[33] I. D. Dinov, K. Siegrist, D. K. Pearl, A. Kalinin, and N. Christou. Probability distributome: a web computational infrastructure for exploring the properties, interrelations, and applications of probability distributions. *Computational statistics*, 31(2):559–577, 2016.

[34] D. Dreger and B. Woods. Regional distance seismic moment tensors of nuclear explosions. *Tectonophysics*, 356(1-3):139–156, 2002.

[35] H. Dufumier and L. Rivera. On the resolution of the isotropic component in moment tensor inversion. *Geophysical Journal International*, 131(3):595–606, 1997.

[36] A. Duijndam. Bayesian estimation in seismic inversion. part i: principles 1. *Geophysical Prospecting*, 36(8):878–898, 1988.

[37] M. M. Dunlop and Y. Yang. Stability of gibbs posteriors from the wasserstein loss for bayesian full waveform inversion. *arXiv preprint arXiv:2004.03730*, 2020.

[38] A. M. Dziewonski, T. Chou, and J. H. Woodhouse. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research: Solid Earth*, 86(B4):2825–2852, 1981.

[39] G. Ekström and M. Nettles. Calibration of the hglp seismograph network and centroid-moment tensor analysis of significant earthquakes of 1976. *Physics of the Earth and Planetary Interiors*, 101(3-4):219–243, 1997.

[40] G. Ekström, M. Nettles, and A. Dziewoński. The global cmt project 2004–2010: Centroid-moment tensors for 13,017 earthquakes. *Physics of the Earth and Planetary Interiors*, 200:1–9, 2012.

[41] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval. Compressed sensing with unknown sensor permutation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1040–1044. Ieee, 2014.

[42] B. Engquist and B. D. Froese. Application of the wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(1):979–988, 2014.

[43] B. Engquist and Y. Yang. Optimal transport based seismic inversion: Beyond cycle skipping. *ArXiv*, abs/2002.00031, 2020.

[44] W. Feller. An introduction to probability theory and its applications. page 99, 1957.

[45] K. Fuchs and G. Müller. Computation of synthetic seismograms with the reflectivity method and comparison with observations. *Geophysical Journal International*, 23(4):417–433, 1971.

[46] S. Fuxian and T. Nafi. Full-waveform based complete moment tensor inversion and source parameter estimation from downhole microseismic data for hydrofracture monitoring. *Geophysics*, 76(6):Wc103–wc116, November 2011. doi: 10.1190/geo2011-0027.1.

[47] O. Gauthier, J. Virieux, and A. Tarantola. Two-dimensional nonlinear inversion of seismic waveforms: Numerical results. *Geophysics*, 51(7):1387–1403, 1986.

[48] L. Gebraad, C. Boehm, and A. Fichtner. Bayesian elastic full-waveform inversion using hamiltonian monte carlo. *Journal of Geophysical Research: Solid Earth*, 125(3):e2019JB018428, 2020.

[49] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[50] S. Gharghabi, S. Imani, A. Bagnall, and E. Keogh. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. In *IEEE Int. Conf. on Data Mining (ICDM2018)*, 2018.

[51] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[52] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[53] W. P. Gouveia and J. A. Scales. Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis. *Journal of Geophysical Research: Solid Earth*, 103(B2):2759–2779, 1998.

[54] D. Grana and E. Della Rossa. Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. *Geophysics*, 75(3):O21–o37, 2010. doi: 10.1190/1.3386676.

[55] P. Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 195:47–63, 2018. doi: 10.1016/j.jspi.2017.09.014.

[56] C. Gu, Y. M. Marzouk, and M. N. Toksöz. Waveform-based bayesian full moment tensor inversion and uncertainty determination for the induced seismicity in an oil/gas field. *Geophysical Journal International*, 212(3):1963–1985, 2017.

[57] L. Guasch, M. Warner, T. Nangoo, J. Morgan, A. Umpleby, I. Stekl, and N. Shah. Elastic 3d full-waveform inversion. In *2012 SEG Annual Meeting*. OnePetro, 2012.

[58] L. Guasch, M. Warner, and C. Ravaut. Adaptive waveform inversion: Practice. *Geophysics*, 84(3):R447–r461, 2019.

[59] A. Guilhem, L. Hutchings, D. S. Dreger, and L. Johnson. Moment tensor inversions of m˜ 3 earthquakes in the geysers geothermal fields, california. *Journal of Geophysical Research: Solid Earth*, 119(3):2121–2137, 2014.

[60] J. Gunning and M. E. Glinsky. Detection of reservoir quality using bayesian seismic inversion. *Geophysics*, 72(3):R37–r49, 2007. doi: 10.1190/1.2713043.

[61] L. Gurobi Optimization. Gurobi optimizer reference manual, 2021.

[62] H. Haario, E. Saksman, J. Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[63] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[64] L. S. House, M. Fehler, J. Barhen, F. Aminzadeh, and S. Larsen. A national laboratory industry collaboration to use seg/eaeg model data sets. *The Leading Edge*, 15(2):135–136, 1996.

[65] L. T. Hu and P. M. Bentler. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods*, 3(4):424, 1998.

[66] W. C. Huang, E. A. Okal, G. Ekström, and M. P. Salganik. Centroid moment tensor solutions for deep earthquakes predating the digital era: the world-wide standardized seismograph network dataset (1962–1976). *Physics of the earth and planetary interiors*, 99(1-2):121–129, 1997.

[67] M. Izzatullah, T. van Leeuwen, and D. Peter. Bayesian uncertainty estimation for full waveform inversion: A numerical study. In *SEG International Exposition and Annual Meeting*. OnePetro, 2019.

[68] W. Jeong, H.-Y. Lee, and D.-J. Min. Full waveform inversion strategy for density in the frequency domain. *Geophysical Journal International*, 188(3):1221–1242, 2012.

[69] B. R. Julian, A. D. Miller, and G. Foulger. Non-double-couple earthquakes 1. theory. *Reviews of Geophysics*, 36(4):525–549, 1998.

[70] J. Kahn, N. Linial, and A. Samorodnitsky. Inclusion-exclusion: Exact and approximate. *Combinatorica*, 16(4):465–477, 1996.

[71] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. doi: 10.1214/12-EJS675.

[72] B. Kleijn, A. Van der Vaart, et al. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012. doi: 10.1214/12-EJS675.

[73] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde. Transport-based analysis, modeling, and learning from signal and data distributions. *arXiv preprint arXiv:1609.04767*, 2016.

[74] D. Komatitsch and J.-P. Vilotte. The spectral element method: an efficient tool to simulate the seismic response of 2d and 3d geological structures. *Bulletin of the seismological society of America*, 88(2):368–392, 1998.

[75] M. Kristeková, J. Kristek, and P. Moczo. Time-frequency misfit and goodness-of-fit criteria for quantitative comparison of time signals. *Geophysical Journal International*, 178(2):813–825, 2009.

[76] O. Lauwers and B. De Moor. A time series distance measure for efficient clustering of input/output signals by their underlying dynamics. *IEEE Control Systems Letters*, 1(2):286–291, 2017. doi: 10.1109/LCSYS.2017.2715399.

[77] X. Li, A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann. Fast randomized full-waveform inversion with compressive sensing. *Geophysics*, 77(3):A13–a17, 2012.

[78] Y. Li, R. Brossier, and L. Métivier. 3d frequency-domain elastic wave modeling with the spectral element method using a massively parallel direct solver. *Geophysics*, 85(2):T71–t88, 2020.

[79] J. Luo, R. S. Wu, and F. Gao. Time-domain full waveform inversion using instantaneous phase information with damping. *Journal of Geophysics and Engineering*, 15(3):1032–1041, 2018.

[80] S. Luo and P. Sava. A deconvolution-based objective function for wave-equation inversion. In *SEG Technical Program Expanded Abstracts 2011*, pages 2788–2792. Society of Exploration Geophysicists, 2011.

[81] B. M, M. T, and G. E. Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: A review. *Geophysics*, 75(5):75a165–75a176, September-October 2010. doi: 10.1190/1.3478209.

[82] Y. Ma and D. Hale. Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion. *Geophysics*, 78(6):R223–r233, 2013.

[83] Y. Ma, D. Hale, B. Gong, and Z. Meng. Image-guided sparse-model full waveform inversion. *Geophysics*, 77(4):R189–r198, 2012.

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

[85] R. Mazumder and H. Wang. Linear regression with mismatched data: A provably optimal local search algorithm. In *Integer Programming and Combinatorial Optimization: 22nd International Conference, IPCO 2021, Atlanta, GA, USA, May 19–21, 2021, Proceedings 22*, pages 443–457. Springer, 2021.

[86] R. Mazumder and H. Wang. Linear regression with partially mismatched data: local search with theoretical guarantees. *arXiv preprint arXiv:2106.02175*, 2021.

[87] L. Métivier, A. Allain, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. A graph-space approach to optimal transport for full waveform inversion. In *SEG Technical Program Expanded Abstracts 2018*, pages 1158–1162. Society of Exploration Geophysicists, 2018.

[88] L. Métivier, R. Brossier, Q. Merigot, and E. Oudet. A graph space optimal transport distance as a generalization of l p distances: application to a seismic imaging inverse problem. *Inverse Problems*, 35(8):085001, 2019.

[89] L. Métivier, R. Brossier, Q. Merigot, E. Oudet, and J. Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysics Journal International*, 205(6):345–377, January 2016. doi: 10.1093/gji/ggw014.

[90] L. Métivier, R. Brossier, Q. Merigot, E. Oudet, and J. Virieux. An optimal transport approach for seismic tomography: Application to 3d full waveform inversion. *IOPscience: Inverse Problems*, 32(11):R59–r80, September 2016. doi: 10.1190/GEO2012-0338.1.

[91] L. Métivier, R. Brossier, J. Virieux, and S. Operto. Full waveform inversion and the truncated newton method. *SIAM Journal of Scientific Computing*, 35(2):B401–b437, July 2013. doi: 10.1137/120877854.

[92] W. Michael, R. Andrew, N. Tenice, M. Joanna, U. Adrian, S. Nikhil, V. Vetle, S. Ivan, G. Lluis, W. Caroline, C. Graham, and A. Bertrand. Anisotropic 3d full-waveform inversion. *Geophysics*, 78(2):R59–r80, March-April 2013. doi: 10.1190/GEO2012-0338.1.

[93] J. W. Miller and D. B. Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, (just-accepted):1–31, 2018. doi: 10.1080/01621459.2018.1469995.

[94] M. Motamed and D. Appelo. Wasserstein metric-driven bayesian inversion with application to wave propagation problems. *arXiv preprint arXiv:1807.09682*, 2018.

[95] M. Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[96] U. K. Müller. Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013. doi: 10.3982/ECTA9097.

[97] H. Neudecker and A. Satorra. A theorem on the rank of a product of matrices. 2009.

[98] G. Odile, V. Jean, and T. Albert. Two-dimensional nonlinear inversion of seismic waveforms: Numerical results. *Geophysics*, 51(7):1387–1403, July 1986.

[99] H. Owhadi, C. Scovel, and T. Sullivan. Brittleness of bayesian inference under finite information in a continuous world. *Electronic Journal of Statistics*, 9(1):1–79, 2015.

[100] A. Pananjady, M. J. Wainwright, and T. A. Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64(5):3286–3300, 2017.

[101] G. Peyré, M. Cuturi, et al. Computational optimal transport. Technical report, 2017.

[102] G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[103] A. Pladys, R. Brossier, Y. Li, and L. Métivier. On cycle-skipping and misfit functions modification for full-wave inversion: comparison of five recent approaches. *Geophysics*, 86(4):1–85, 2021.

[104] R. E. Plessix. Three-dimensional frequency-domain full-waveform inversion with an iterative solver. *Geophysics*, 74(6):Wcc149–wcc157, November-December 2009. doi: 10.1190/1.3211198.

[105] A. Postnikov. Permutohedra, associahedra, and beyond. *International Mathematics Research Notices*, 2009(6):1026–1106, 2009.

[106] J. Ramsay. Functional data analysis. *Encyclopedia of Statistics in Behavioral Science*, 2005.

[107] A. Ray, S. Kaplan, J. Washbourne, and U. Albertin. Low frequency full waveform seismic inversion within a tree based bayesian framework. *Geophysical Journal International*, 212(1):522–542, 2017.

[108] A. Ray, A. Sekar, G. M. Hoversten, and U. Albertin. Frequency domain full waveform elastic inversion of marine seismic data from the alba field using a bayesian trans-dimensional algorithm. *Geophysical Journal International*, 205(2):915–937, 2016. doi: 10.1093/gji/ggw061.

[109] F. Rickers, A. Fichtner, and J. Trampert. Imaging mantle plumes with instantaneous phase measurements of diffracted waves. *Geophysical Journal International*, 190(1):650–664, 2012.

[110] B. Romain, O. Stephane, and V. Jean. Seismic imaging of complex onshore structures by 2d elastic frequency-domain full-waveform inversion. *Geophysics*, 74(6):Wcc105–wcc118, November-December 2009. doi: 0.1190/1.3215771.

[111] H. S. Sánchez-Reyes, J. Tago, L. Métivier, V. Cruz-Atienza, and J. Virieux. An evolutive linear kinematic source inversion. *Journal of Geophysical Research: Solid Earth*, 123(6):4859–4885, 2018.

[112] A. Scarinci, M. Fehler, and Y. Marzouk. Robust bayesian moment tensor inversion using transport-lagrangian distances. In *SEG Technical Program Expanded Abstracts 2019*, pages 2123–2127. Society of Exploration Geophysicists, 2019.

[113] F. Schorfheide. Loss function-based evaluation of dsge models. *Journal of Applied Econometrics*, 15(6):645–670, 2000.

[114] B. Schwartz. A computational analysis of the auction algorithm. *European journal of operational research*, 74(1):161–169, 1994.

[115] M. K. Sen and P. L. Stoffa. Bayesian inference, gibbs' sampler and uncertainty estimation in geophysical inversion. *Geophysical Prospecting*, 44(2):313–350, 1996.

[116] P. M. Shearer. *Introduction to seismology*. Cambridge University Press, 2009.

[117] M. K. Simon. *Probability distributions involving Gaussian random variables: A handbook for engineers and scientists*. Springer Science & Business Media, 2007.

[118] S. Singh, I. Tsvankin, and E. Z. Naeini. Bayesian framework for elastic full-waveform inversion with facies information. *The Leading Edge*, 37(12):924–931, 2018.

[119] R. C. Smith. *Uncertainty quantification: theory, implementation, and applications*, volume 12. Siam, 2013.

[120] L. V. Socco, S. Foti, and D. Boiero. Surface-wave analysis for building near-surface velocity models—established approaches and new perspectives. *Geophysics*, 75(5):75a83–75a102, 2010.

[121] K. Spikes, T. Mukerji, J. Dvorkin, and G. Mavko. Probabilistic seismic inversion based on rock-physics models. *Geophysics*, 72(5):R87–r97, 2007.

[122] B. Stephen, H. Lior, A. Aleksandr, and Z. Sergiy. General optimization framework for robust and regularized 3d full waveform inversion. In *77th European Association of Geoscientists and Engineers (EAGE) Conference and Exhibition*, Madrid, Spain, June 2015. Ifema.

[123] A. Stovas and Y. Roganov. *Acoustic waves in layered media-from theory to seismic applications*. IntechOpen London, 2011.

[124] M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.

[125] W. W. Symes. Migration velocity analysis and waveform inversion. *Geophysical prospecting*, 56(6):765–790, 2008.

[126] S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

[127] B.-T. Tan, G. Omar, M. James, and S. Georg. A computational framework for infinite-dimensional bayesian inverse problems part i: the linearized case, with application to global seismic inversion. *SIAM Journal of Scientific Computing*, 35(6):A2494–a2523, September-October 2013. doi: 10.1137/12089586X.

[128] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepcev. A transportation l_p distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2):187–210, 2017.

[129] M. Thorpe and D. Slepcev. Transportation l_p distances: Properties and extensions, 2017.

[130] J. Thurin, R. Brossier, and L. Métivier. An ensemble-transform kalman filter: Full-waveform inversion scheme for uncertainty estimation. In *SEG Technical Program Expanded Abstracts 2017*, pages 1307–1313. Society of Exploration Geophysicists, 2017.

[131] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[132] V. L. Tristan and H. Felix. Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195(2):661–667, June 2013. doi: 10.1093/gji/ggt258.

[133] B. Ursin. Review of elastic and electromagnetic wave propagation in horizontally layered media. *Geophysics*, 48(8):1063–1081, 1983.

[134] V. Vaclav and K. Daniela. Moment tensor inversion of waveforms: a two-step time-frequency approach. *Geophysical Journal International*, 190:1761–1776, June 2012. doi: 10.1111/j.1365-246X.2012.05592.x.

[135] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[136] T. Van Leeuwen and W. Mulder. A correlation-based misfit criterion for wave-equation traveltime tomography. *Geophysical Journal International*, 182(3):1383–1394, 2010.

[137] V. Vavryčuk. Moment tensor decompositions revisited. *Journal of Seismology*, 19(1):231–252, 2015.

[138] D. Vigh, B. Starr, J. Kapoor, and H. Li. 3d full waveform inversion on a gulf of mexico waz data set. In *SEG Technical Program Expanded Abstracts 2010*, pages 957–961. Society of Exploration Geophysicists, 2010.

[139] D. Vigh and E. W. Starr. 3d prestack plane-wave, full-waveform inversion. *Geophysics*, 73(5):Ve135–ve144, 2008.

[140] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[141] J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):Wcc1–wcc26, 2009.

[142] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[143] T. C. Wallace. A reexamination of the moment tensor solutions of the 1980 mammoth lakes earthquakes. *Journal of Geophysical Research: Solid Earth*, 90(B13):11171–11176, 1985.

[144] M. Warner and L. Guasch. Adaptive waveform inversion-fwi without cycle skipping-theory. In *76th EAGE Conference and Exhibition 2014*, volume 2014, pages 1–5. European Association of Geoscientists & Engineers, 2014.

[145] M. Warner and L. Guasch. Adaptive waveform inversion: Theory. *Geophysics*, 81(6):R429–r445, 2016.

[146] M. Warner, T. Nangoo, N. Shah, A. Umpleby, and J. Morgan. Full-waveform inversion of cycle-skipped seismic data by frequency down-shifting. In *SEG Technical Program Expanded Abstracts 2013*, pages 903–907. Society of Exploration Geophysicists, 2013.

[147] M. Warner, A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Štekl, L. Guasch, C. Win, et al. Anisotropic 3d full-waveform inversion. *Geophysics*, 78(2):R59–r80, 2013.

[148] J. Watson, C. Holmes, et al. Approximate models and robust decisions. *Statistical Science*, 31(4):465–489, 2016. doi: 10.1214/16-STS592.

[149] H. White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[150] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[151] L. A. Wolsey. *Integer programming*. John Wiley & Sons, 2020.

[152] L. Xiang, A. Aleksandr, v. L. Tristan, and H. Felix. Fast randomized full-waveform inversion with compressive sensing. *Geophysics*, 77(3):A13–a17, May-June 2012. doi: 10.1190/GEO2011-0410.1.

[153] F. Yang and J. Ma. Deep-learning inversion: A next-generation seismic velocity model building method. *Geophysics*, 84(4):R583–r599, 2019.

[154] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt. Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–r62, 2018.

[155] H. Yuan, S. French, P. Cupillard, and B. Romanowicz. Lithospheric expression of geological units in central and eastern north america from full waveform tomography. *Earth and Planetary Science Letters*, 402:176–186, 2014.

[156] E. L. Yunyue and D. Laurent. Full-waveform inversion with extrapolated low-frequency data. *Geophysics*, 81(6):R339–r348, November-December 2016. doi: 10.1190/GEO2016-0038.1.

[157] H. Zhu, S. Li, S. Fomel, G. Stadler, and O. Ghattas. A bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration. *Geophysics*, 81(5):R307–r323, 2016.

[158] L. Zhu, E. Liu, and J. H. McClellan. Sparse-promoting full-waveform inversion based on online orthonormal dictionary learning. *Geophysics*, 82(2):R87–r107, 2017.