Investigations on the discrimination ability of multivariate scoring rules

DISSERTATION

zur Erlangung des Grades eines Doktors
der Naturwissenschaften

vorgelegt von

Maximilian Stock

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät
der Universität Siegen

Siegen 2022

Betreuer und erster Gutachter

Prof. Dr. Alfred Müller

Universität Siegen

Zweiter Gutachter

Jun.-Prof. Dr. Marco Oesting

Universität Stuttgart

Tag der mündlichen Prüfung:

22. September 2022

Abstract

Probabilistic forecasts in form of predictive distributions over future quantities have become more and more important in many different fields, including meteorology, hydrology, epidemiology and economics.
Along with the growing prevalence of probabilistic models the need for tools to evaluate the appropriateness of models and forecasts emerges. Various measures have been developed to address this topic.
In their seminal paper, Gneiting and Raftery (2007) study so called *proper scoring rules* as summary measures to evaluate probabilistic forecasts by assigning a single numerical score based on the predictive distribution and the event that materializes. Such a proper scoring rule encourages the forecaster to make careful assessments.
For univariate quantities there is a vast selection of proper scoring rules and their properties are understood quite well. However, for multivariate quantities there is only a small number of proper scoring rules available and there does not yet exist much research on their properties. The most prominent example of a strictly proper multivariate scoring rules is the *energy score*.
A scoring rule should not only be strictly proper, but it should also assign significantly different score values to probabilistic forecasts of models that are significantly wrong. This property is refered to as discrimination ability of a scoring rule.
To assess the discrimination ability of a scoring rule we use the Diebold-Mariano test, which is a crucial element in the evaluation of scoring rules. In this thesis we mainly focus on the discrimination ability of scoring rules for multivariate distributions with a special emphasis on the correct modelling of the dependence structure. The energy score has been criticized for its poor ability to distinguish between forecasting distributions with different dependence structure, whereas it detects very well errors in location and scale.
The discrimination ability of the energy score depends on the choice of the parameter β, which is very often fixed to 1. However, β can be chosen to be any value in the interval $(0, 2)$. Thus, the main topic of this thesis is to study the discrimination ability of the energy score for various choices of the parameter β and to compare it with other known scoring rules like the Dawid-Sebastiani score and the variogram score. An extensive simulation study shows that the discrimination ability of the energy score typically improves with smaller parameter β. Therefore, a new multivariate strictly proper scoring rule is introduced that arises as a scaling limit of the energy score as β tends to zero and its properties are studied.

Zusammenfassung

Probabilistische Vorhersagen über zukünftige Größen in Form von Wahrscheinlichkeitsverteilungen haben in vielen verschiedenen Bereichen, einschließlich der Metereologie, der Hydrologie, der Epidemiologie und den Wirtschaftswissenschaften, immer mehr an Bedeutung gewonnen.
Mit der zunehmenden Verbreitung probabilistischer Modelle wächst auch der Bedarf an Methoden zur Bewertung der Angemessenheit von Modellen und Prognosen. Es wurden diverse Maße entwickelt, um dieses Thema anzugehen.
In ihrer bahnbrechenden Arbeit untersuchen Gneiting and Raftery (2007) so genannte *proper scoring rules* als Maße zur Bewertung probabilistischer Vorhersagen, indem sie der vorhergesagten Verteilung und einem eingetretenen Ereignis einen score zuweisen. Eine solche proper scoring rule bietet dem Ersteller der Vorhersage einen Anreiz, sorgfältige Bewertungen vorzunehmen.
Für univariate Größen gibt es eine große Auswahl an geeigneten proper scoring rules und ihre Eigenschaften sind recht gut bekannt. Für multivariate Größen gibt es jedoch nur eine kleine Anzahl geeigneter proper scoring rules. Auch sind deren Eigenschaften bisher noch nicht ausreichend erforscht. Das bekannteste Beispiel für eine strikte proper scoring rule ist der *energy score*.
Eine scoring rule sollte allerdings nicht nur strikt proper sein, sondern auch probabilsitischen Vorhersagen von Modellen, die signifikant falsch sind, signifikant unterschiedliche Werte zuweisen. Diese Eigenschaft wird als Fähigkeit zur Diskriminierung einer scoring rule bezeichnet.
Um diese Unterscheidungsfähigkeit zu beurteilen, verwenden wir den Diebold-Mariano Test, der ein entscheidendes Element bei der Bewertung von scoring rules ist. In dieser Arbeit konzentrieren wir uns hauptsächlich auf die Unterscheidungsfähigkeit von scoring rules für multivariate Verteilungen mit besonderem Augenmerk auf die korrekte Modellierung der Abhängigkeitsstruktur. Der energy score wurde oft kritisiert, weil er nicht in der Lage ist, zwischen Verteilungen mit unterschiedlicher Abhängigkeitsstruktur zu unterscheiden, während er Fehler der Lokations- und Skalenparameter sehr gut erkennt.
Die Unterscheidungsfähigkeit des energy scores hängt von der Wahl des Parameters β ab, der häufig auf 1 festgelegt ist. Jedoch kann für β jeder beliebige Wert im Intervall $(0, 2)$ gewählt werden. Das Hauptthema dieser Arbeit ist daher die Untersuchung der Unterscheidungsfähigkeit des energy scores für verschiedene Auswahlmöglichkeiten des Parameters β und der Vergleich mit anderen bekannten scoring rules wie dem *Dawid-Sebastiani score* und dem *variogram score*. Eine umfangreiche Simulationsstudie zeigt, dass sich die Unterscheidungsfähigkeit des energy scores typischerweise mit kleinerem Parameter β verbessert. Im Zuge dessen wird eine neue multivariate strikte proper scoring rule eingeführt, die sich als skalierter Grenzwert des energy scores ergibt, wenn β gegen Null geht, und ihre Eigenschaften werden untersucht.

Contents

1. Introduction

A key desire of mankind is making forecasts about an uncertain future. As any prediction is typically surrounded by uncertainty, forecasts should be probabilistic in nature, taking the form of probability distributions over future quantities or events, see Dawid (1984). Probabilistic forecasts allow to quantify the inherent uncertainty of a forecast and thus serve for good decision making.
Accordingly, probabilistic forecasts in form of predictive distributions over future quantities or events have become popular over the last few decades in many different fields, including meteorology, hydrology, seismology, economics, finance, demographic and political science.
Prominent examples among others are the inflation reports iussed by the Bank of England, see e.g. Bailey (2021), and the use of ensemble forecasts in meteorology, see Gneiting, Raftery, et al. (2005), and Leutbecher and Palmer (2008). For a recent and extensive review of probabilistic forecasts we refer to Gneiting and Katzfuss (2014).
Along with the growing prevalence of probabilistic models the need for tools to evaluate the appropriateness of models and forecasts emerges. Over the past years various measures have been developed to address this demand.
Proper scoring rules provide summary measures to evaluate probabilistic forecasts by assigning a single numerical score based on the predictive distribution and the event that materializes. A scoring rule is proper, if the forecaster minimizes the expected score for an observation drawn from distribution F if he or she issues the probabilistic forecast F, rather than $G \neq F$. The scoring rule is *strictly proper*, if the minimum is unique. Therefore, in prediction problems scoring rules encourage the forecaster to make careful assessments.
While there exists a vast selection of proper and strictly proper scoring rules for univariate quantities, there is only a very limited amount of scoring rules for multivariate quantities available and many of them require that the forecast is given in form of a probability density function.
The most commonly utilized multivariate strictly proper scoring rule is the *energy score*, which draws back on Székely's energy distance and can be seen as a multivariate extension of the univariate continuous ranked probability score, see Székely (2003). The energy score is also readily applicable to the important case of ensemble forecasts. However, the energy score is frequently criticized in literature due to its apparently poor ability to detect incorrectly specified correlations between the components of the multivariate quantity, see for instance Pinson and Tastu (2013), and Scheuerer and Hamill (2015).
The main focus of the work presented in this thesis will be the discrimination ability

of the energy score. The energy score depends on a parameter β which in literature is essentially fixed to 1. However, β can take any value in the interval $(0, 2)$, so we study the discrimination ability of the energy score for various parameters.
Furthermore, the poor discrimination ability of the energy score attributed in literature draws back on the working paper of Pinson and Tastu (2013), in which the discrimination ability of the energy score is merely evaluated with the relative change in score. In this thesis we will argue that this metric is not sufficient. We introduce the Diebold-Mariano test which is a crucial element in score evaluation.
With these considerations in mind we are able to define a new strictly proper scoring rule for multivariate distributions that arises as a limiting case of the energy score as β tends to zero.

The structure of this thesis is as follows. In Chapter 2 we introduce the term probabilistic forecast. A probabilistic forecast takes the form of a predictive distribution over future quantities or events of interest. Probabilistic forecasting aims to maximize the sharpness of the predictive distribution, subject to calibration, see Gneiting and Katzfuss (2014).

In Chapter 3 we define the term (strictly) proper scoring rule. Proper scoring rules provide summary measures to assess both calibration and sharpness of a probabilistic forecast by assigning a single numerical value to a forecasting distribution based on an event that materializes. We also introduce the most common scoring rules for multivariate quantities, namely the energy score, the variogram score, and the Dawid-Sebastiani score.

After considering the theoretical properties of probabilistic forecasts and proper scoring rules, in Chapter 4 we deal with the reporting of probabilistic forecasts in practical application. There are essentially two reporting options, namely reporting the forecasting distribution in form of its CDF or in form of an ensemble forecast. The second reporting option is commonly used in practical application, particularly in weather and climate prediction.

In Chapter 5 we introduce estimators for the score values corresponding to the scoring rules introduced in Chapter 3. Note that we have to estimate the score values based on an event that materializes and the forecasting distribution which is usually available in form of an ensemble forecast. Particularly, it holds that the true distribution which we aim to forecast generally is unkown in practice.

Chapter 6 deals with the evaluation of the predictive performance of a probabilistic forecast. There are different measures for evaluation of the score values corresponding to different probabilistic forecasts. The most important criterion is the Diebold-Mariano test which is a test to determine whether two forecasts are significantly different.

Afterwards we introduce the concept of discrimination ability of a proper scoring rule in Chapter 7. This is a central concept in this thesis and refers to the ability of a scoring rule to discriminate between different forecasting distributions. If a scoring rule discriminates well between different probabilistic forecasts, the resulting score values should also differ significantly. A crucial tool for the determination of the discrimination ability of a scoring rule is the Diebold-Mariano test.
In this chapter we execute several simulation studies to consider the discrimination ability of the energy score, the variogram score and the Dawid-Sebastiani score with repect to miscalibrated forecasting distributions. Our main focus here is to study the discrimination ability of the energy score with different parameters β for forecasting distributions that differ in their interdependence structure.

As the energy score is the scoring rule that overall has the best discrimination ability of the considered scoring rules, we consider the related energy distance more closely in Chapter 8.

Afterwards, we consider the limiting case of the energy score as β tends to zero in Chapter 9. In the course of this we are able to define a new multivariate scoring rule that is strictly proper relative to a broad class of distributions, so this scoring rule is useful in application.

Finally, Chapter 10 summarizes the results of this thesis and gives a short overview on future research questions.

2. Probabilistic forecasting

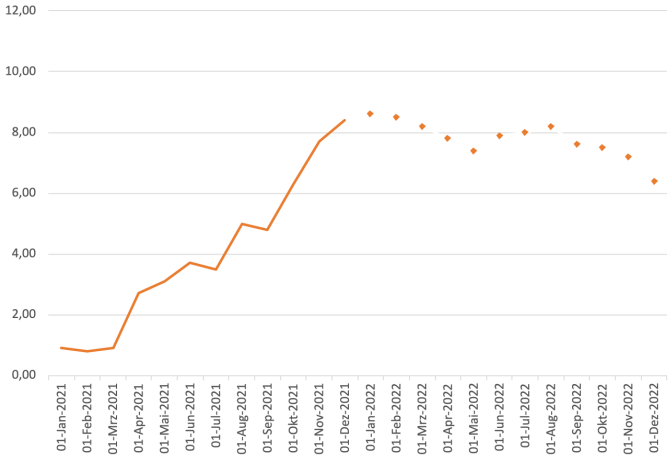
It is a great human desire to make predictions about the future. As the future naturally is uncertain, forecasts should be probabilistic in nature taking the form of probability distributions over future quantities or events, see Gneiting and Katzfuss (2014). In the past few decades probabilistic forecasts have become more and more popular in various fields and one could witness a paradigm shift from single-valued or point forecasting to distributional or probabilistic forecasting in various key applications, as reviewed by Gneiting and Katzfuss (2014), Tay and Wallis (2000), Gneiting, Stanberry, et al. (2008), Timmermann (2000), and many others. The transition from point to distributional prediction is described in Stigler (1975) from a historical perspective.

2.1. Preliminaries on probabilistic forecasting

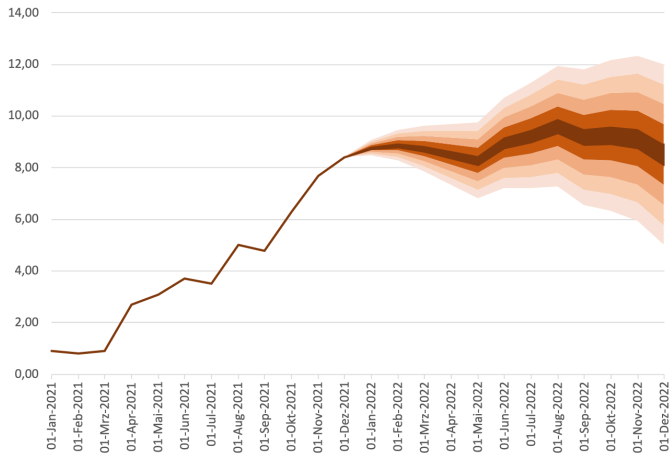
The term point forecasting denotes a forecasting method, where the future is merely associated with a single expected outcome which is usually an average expected value. In contrast, probabilistic forecasts assess the uncertainty associated with the forecast by allocating a probability for different events to happen.
Probabilistic forecasts, therefore, provide a complete description of the uncertainty associated with a prediction in contrast to a point forecast which by itself contains no description of the associated uncertainty, see Tay and Wallis (2000).

An easy example for both concepts is rolling a dice: No one will forecast the outcome as 3.5, which corresponds to point forecasting the expected value. Instead, we understand that we have a chance of $1/6$ to get each of the numbers from 1 to 6.

Since we live in an uncertain world, probabilistic forecasts should be preferred to point forecasts as they serve for optimal decision making by inherently assessing the uncertainty associated with the forecast. While probabilistic forecasts for binary events have been commonly iussed for several decades, today attention is shifting towards more general types of events, including multicategorial and continuous variables.



(a) Point forecast



(b) Probabilistic forecast

Figure 2.1.: Illustration of a point forecast and a probabilistic forecast via hypothetical inflation rates. The line chart in both figures represents the observed historical data. In the left figure we see a point forecast in form of the expectation values illustrated by the individual dots. In the right figure the probabilistic forecast is represented in form of a fan chart. The central area represents a pointwise 10 % prediction interval. The lighter shaded areas extend this interval each by 10 %, so that the inflation rate is expected somewhere within these fans on 90 out of 100 occasions.

A key application of probabilistic forecasting is weather and climate prediction, see e.g. Gneiting, Raftery, et al. (2005), Palmer (2002), and Gel, Raftery, and Gneiting (2004). One of the simplest examples is the issuing about rainfall in the form of the probability of precipation, see Bianco (2021).
Other popular fields are hydrologic forecasting including flood risk assessment, see Cloke and Pappenberger (2009), and also seismic hazard prediction, see T. Jordan et al. (2011).
Probabilistic forecasting is also becoming increasingly popular in energy forecasting, particularly in forecasting the availability of renewable energy sources such as wind and solar power, see Pinson (2013). Other important fields are economic and financial risk management, see Timmermann (2000).

Example 2.1.1. As an illustrating example we consider the probabilistic forecasts of inflation rates iussed by the Bank of England's Monetary Policy Commitee for about two decades. Figure 2.2 is taken from the November 2021 report, see Bailey (2021). It is a projection of the future UK consumer price index (CPI). The fan chart[1] shows the predictive distribution in terms of annual percentage change. The central area represents a pointwise 30 % prediction interval, and the lighter-shaded bands extend this interval each by 30 %. The inflation is therefore expected to lie within the fans on 90 out of 100 occasions. On the remaining 10 out of 100 occasions inflation can fall outside the red area of the fan chart. Over the forecast period this is depicted as the light grey background.

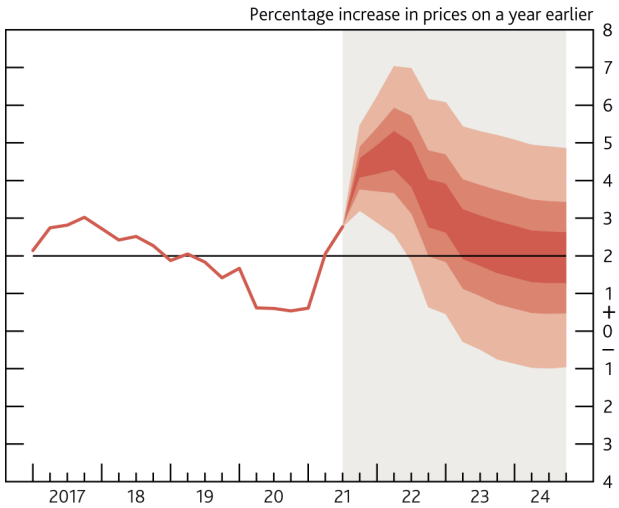


Figure 2.2.: November 2021 Bank of England forecast of inflation in the United Kingdom as a percentage increase in the consumer price index in percent for the following years.

[1]The term fan chart was coined by the Bank of England which has been using these charts and this term since 1997 in its inflation report to describe its forecasted inflation to the general public. It is a commonly used tool in finance to visualize probabilistic forecasts, see Britton, Fisher, and Whitley (1998).

2.2. Mathematical framework

The concept of probabilistic forecasts can be formalized as described in the following section. In their seminal paper Murphy and Winkler (1987) called for the consideration of the joint distribution of the forecast and the observation. While their work was focused on the setting of point forecasts, this framework was adapted and extended by Gneiting and Ranjan (2013), Ehm, Gneiting, et al. (2016), and Strähl and Ziegel (2017) to include the case of potentially multiple probabilistic forecasts.
In the following, we will introduce the key tool of a prediction space, and the notions of calibration and sharpness. Furthermore, we will argue that probabilistic forecasting should aim to maximize the sharpness of the predictive distribution subject to calibration, see Murphy and Winkler (1987), and Gneiting and Raftery (2007).
The general setting considers the joint distribution of forecasts and observations on a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$, where the elements of the sample space Ω can be identified with tuples

$$(F_1, \dots, F_k, Y),$$

whose distribution is specified by the probability measure $\mathbb{Q}$. The probability distributions $F_1, \dots, F_k$ are probability measures on the outcome space $(\Omega_Y, \mathcal{A}_Y)$ for the observation Y.
In this work we restrict our attention to real-valued observations, i.e. $\Omega_Y = \mathbb{R}^d$, unless stated otherwise. In this case the probabilistic forecasts F_i can be identified with the associated cumulative distribution function (CDF) F_i or probability density function (PDF) f_i if the distribution is absolutely continuous with respect to Lebesgue measure. Note that this not only includes the case of forecasting a d-dimensional quantity but also the important case of a (fully) d-step ahead forecast of an univariate time series, where the complete d-step ahead distribution is covered. In particular, not only the marginals for every time step are predicted but also the complete dependency structure. Nevertheless, Ω can be any set. This of course also includes discrete sets.

2.3. Calibration and sharpness

In the following, we review the notions of calibration and sharpness, which are basic and initial tools to evaluate probabilisitic forecasts for univariate quantities. So in this section we are concerned with the case of a real-valued variable of interest Y, for which a probabilistic forecast can be identified with the cumulative distribution function F on the real line $\mathbb{R}$.
As argued precisely in Gneiting and Raftery (2007) the general aim of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration.
Calibration is a joint property of the predictive distribution and the associated observations. It concerns the statistical compatibility between the probabilistic forecasts and the realizations. Essentially, the observations should be indistinguishable from random

draws from the predictive distributions, see Gneiting and Katzfuss (2014).
Sharpness is a property of the forecast only. It refers to the concentration of the predictive distribution.
Various notions of calibration have been proposed in literature. For now, we focus on probabilistic calibration. A probabilistic forecast F is probabilistic calibrated if the probability integral transform (PIT) $F(Y)$ is uniformly distributed, with appropriate technical adaptions in cases where F may have a discrete component, see Gneiting and Raftery (2007) and Gneiting and Ranjan (2013).
Given that a probabilistic forecast is calibrated it should be as sharp as possible, as more concentrated forecasting distributions indicate a higher information content in the prediction. In the case of density forecasts for real-valued variables, sharpness can be assessed in terms of the prediction intervals. The mean width of these intervals should be as short as possible, subject to the empirical coverage being at the nominal level, see Gneiting and Katzfuss (2014).
There exist various tools for assessing calibration and sharpness in practical applications. For CDF-valued probabilistic forecasts an essential device is provided by checks of the uniformity of the PIT values. Given a sample of pairs of probabilistic forecasts and observations (F_i, y_i), $i = 1, \ldots, N$ calibration can be assessed by visual inspection of the histogram of the PIT-values $F_i(y_i)$, $i = 1, \ldots, N$, see Dawid (1984), and Gneiting and Raftery (2007). Deviations of the PIT-values from the desired uniform distribution may indicate miscalibration.
The shape of the histogram can point towards the reasons of miscalibration, e.g. an U-shaped histogram indicates an underdispersed forecasting distribution with too narrow prediction intervals, whereas an inverse U-shaped histogram indicates an overdispersed forecasting distribution with too wide prediction intervals.
Checks of calibration via the uniformity of PIT histograms should by accompanied by an assessment of sharpness. Otherwise misspecifications in the forecasting distributions can remain undetected, see Gneiting and Raftery (2007), and Diebold, Hahn, and Tay (1999).
Also formal statistical tests of the hypothesis that a probabilistic forecast is calibrated can be applied, provided that these tests account for typically complex dependence structures, particularly in the case of univariate time series forecasting, see Corradi and Swanson (2006), and Knüppel (2015).
An alternative tool to assess calibration and sharpness is considering the coverage and width of prediction intervals. The coverage of a $(1-\alpha)\cdot 100\,\%$, $\alpha \in (0,1)$, central prediction interval is the proportion of validating observations located between the lower and upper $\alpha/2$ -quantiles of the predictive distribution. For calibrated probabilistic forecasts it should be around $(1-\alpha)\cdot 100\,\%$. Sharper forecasting distributions lead to a narrower central prediction interval, therefore the width of these intervals depicts a natural measure for the sharpness of a probabilistic forecast.

3. Proper scoring rules

Proper scoring rules provide summary measures for assessing both calibration and sharpness by assigning a numerical value based on the predictive distribution F and the event y that materializes. The role of scoring rules in terms of elicitation is to encourage the assessor to make careful assessments. In terms of evaluation the role of scoring rules is to measure the quality of probabilistic forecasts and to rank competing forecasts, see Gneiting and Raftery (2007).
Generally we assume scoring rules to be negatively oriented, i.e. a lower value indicates that the probabilistic forecast has a better quality. Thus, a scoring rule can be viewed as a penality that the forecaster wants to minimize. Some authors assume scoring rules to be positively oriented, see Gneiting and Raftery (2007), so that the score rewards the assessor. However, the following definitions and results can also be applied in this case.

3.1. Notations and basics

As in the preceding section, let Ω_Y be the sample space, that is the set of possible values of the variable of interest, and let $\mathcal{A}_Y$ be a σ-algebra on Ω_Y. Furthermore, let $\mathcal{P}$ be a convex class of probability measures on $(\Omega_Y, \mathcal{A}_Y)$. A probabilistic forecast is any probability measure $F \in \mathcal{P}$.

Definition 3.1.1. A *scoring rule* Sc is defined as an extend real-valued function

$$\mathrm{Sc} : \mathcal{P} \times \Omega_Y \to \overline{\mathbb{R}},$$

such that $\mathrm{Sc}(F, \cdot)$ is $\mathcal{P}$-quasi-integrable for all $F \in \mathcal{P}$.

This means a scoring rule is functional assigning a value to the association of a predictive probability measure F with an observation y from the real probability measure of the random variable. We write

$$\mathrm{Sc}(F, G) = \int \mathrm{Sc}(F, y) \mathrm{d}G(y) \tag{3.1}$$

for the *expected score* under G when the probabilistic forecast is F. Note that G is the distribution which we aim to forecast.

Definition 3.1.2. A scoring rule Sc is called *proper* relative to the class $\mathcal{P}$ if

$$\mathrm{Sc}(G, G) \leq \mathrm{Sc}(F, G) \quad \text{for all } F, G \in \mathcal{P}, \tag{3.2}$$

i.e. the expected score is optimized if the true distribution of the observation is issued as a forecast. It is called *strictly proper* relative to the class $\mathcal{P}$, if (3.2) holds with equality only if $F = G$.

Thus, a scoring rule is designed in a way that quoting the true distribution as the forecasting distribution is an optimal strategy in expectation. This property is crucial, as the use of improper scoring rules can lead to heavily misguided conclusions about the predictive performance of a probabilistic forecast, see Gneiting and Raftery (2007), Gneiting (2011), and Hendrickson and Buehler (1971).
The general idea of propriety dates back at least to Brier (1950), and Good (1952). However, the term "proper" was apparently coined by Winkler and Murphy (1968).

Definition 3.1.3. A scoring rule $\mathrm{Sc} : \Omega_Y \times \mathcal{P} \to \overline{\mathbb{R}}$ is called *regular* relative the class $\mathcal{P}$, if $\mathrm{Sc}(F, G)$ is real-valued for all $F, G \in \mathcal{P}$, expect that we allow $\mathrm{Sc}(F, G) = \infty$ if $F \neq G$.

For a detailed mathematical analysis of properties and characterizations of proper scoring rules we refer to Gneiting and Raftery (2007). In the subsequent section connections to convex analysis are established by measure-theoretic representations, see Rockafellar (1970). For related work on local proper scoring rules we refer to Ehm and Gneiting (2012), Parry, Dawid, and Lauritzen (2012), and Ovcharov (2015). Local proper scoring rules only depend on the forecast density through its value and the value of its derivates at the observation y.

3.2. Score divergences

Directly associated with a given scoring rule Sc is the *score divergence*.

Definition 3.2.1. If the scoring rule Sc is regular and proper we call the non-negative function

$$d(F, G) = \mathrm{Sc}(F, G) - \mathrm{Sc}(G, G) \tag{3.3}$$

the associated *divergence function*, see Gneiting and Raftery (2007).

Clearly, the following holds:

Remark 3.2.2. If the scoring rule Sc is strictly proper, the associated divergence function $d(F, G)$ is strictly positiv, unless $F = G$.

Score divergences are closely related to the concept of Bregman-divergences, see Bregman (1967). The connections can be revealed by representations of proper scoring rules as supergradients of concave functions, see Gneiting and Raftery (2007), and Hendrickson and Buehler (1971).

The result of the following theorem, see Gneiting and Raftery (2007), characterizes proper scoring rules using the tools of convex analysis, see Rockafellar (1970).

Theorem 3.2.3. The scoring rule Sc is proper relative to the class $\mathcal{P}$ if and only if it is regular, and the expected score function (or entropy) $e(F) := \mathrm{Sc}(F, F)$ is concave and $\mathrm{Sc}(F, \cdot)$ is a supergradient of e at the point F for all $F \in \mathcal{P}$.

Example 3.2.4. Let $\mathcal{P}$ denote the class of probability measures with a square-integrable Lebesgue-density f. Consider the quadratic score

$$\mathrm{QS}(F, y) = -2f(y) + \int f^2(z)\mathrm{d}z.$$

Then $e(F)$ is concave with supergradient $\mathrm{QS}(F, \cdot)$. Thus, QS is proper. An interesting observation here is that although the linear score $\mathrm{LS}(F, y) = -f(y)$ has the same expected score function $e(F)$, the linear score is not a supergradient, and, therefore, is improper.

If the sample space is finite and the expected score function is sufficiently smooth, the divergence function becomes the *Bregman divergence*, see Bregman (1967), which plays a major role in optimization and has recently attracted the attention of the machine learning community.
In case of infinite sample spaces, e.g. if $\Omega_Y = \mathbb{R}$, technical modifications such as extensions to functional Bregman divergences are required, see Frigyik, Srivastava, and Gupta (2008), and Ovcharov (2018).
The representations of proper scoring rules and the connection to Bregman divergences reveal the close relation of proper scoring rules and convex analysis.
For real-valued quantities there is a broad range of scoring rules available and the literature is rather sophisticated. However, for multivariate forecasts there is only a very limited amount of scoring rules available. In the following, we will recall the major important ones.

3.3. Continuous ranked probability score

The *continuous ranked probability score* (CRPS) is one of the most widely used scoring rules for $\Omega_Y = \mathbb{R}$. Let $\mathcal{P}$ consist of the Borel probability measures on $\mathbb{R}$. We identify a probabilistic forecast, i.e. a member of the class $\mathcal{P}$ with its cumulative distribution function F.

Definition 3.3.1 (see Matheson and Winkler (1976))**.** Let F be a CDF-valued probabilistic forecast and $y \in \mathbb{R}$ be an observation. The CRPS is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}_{\{y \leq x\}})^2 \mathrm{d}x. \tag{3.4}$$

Remark 3.3.2. The function

$$\mathrm{BS}_x(F, y) = (F(x) - \mathbb{1}_{\{y \leq x\}})^2$$

is a proper scoring rule itself, namely the Brier score, see Brier (1950).

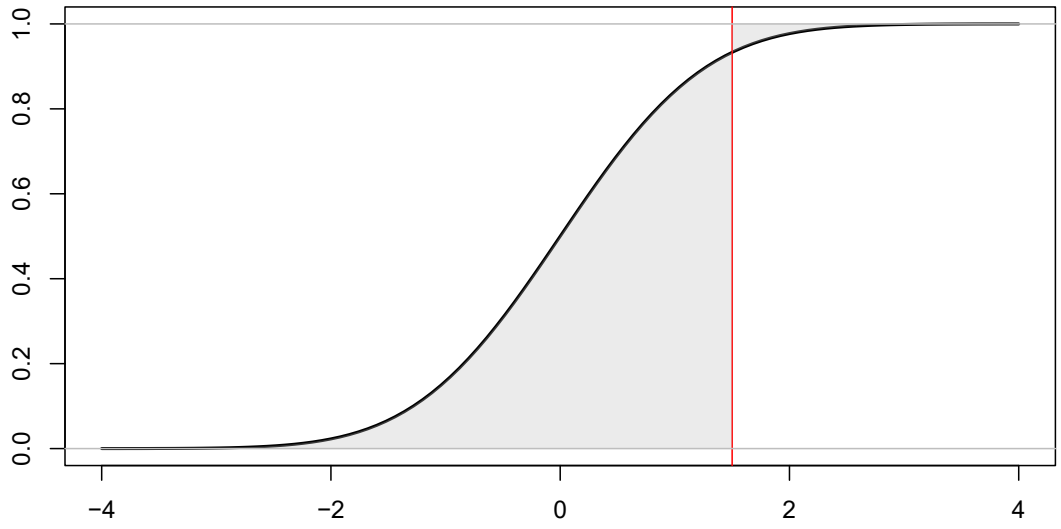


Figure 3.1.: Schematic CRPS. We use $F = \Phi(0, 1)$ the standard normal distribution function, and $y = 1.5$ to illustrate the concept of the CRPS. The forecasting distribution F is penalized for the grey shaded area left and right of the realized value y through $\int_{-\infty}^{y} F(z)^2 \mathrm{d}z$ and $\int_{y}^{\infty} (1 - F(z))^2 \mathrm{d}z$ respectively. A lower score suggests a higher sharpness of the forecasting distribution around the realization y.

If the first moment of F is finite, the CRPS can be represented alternatively. To show this we need to following lemma, see Baringhaus and Franz (2004).

Lemma 3.3.3. Let X and Y be independent real random variables with finite expectations. Let F be the distribution function of X, and G be the distribution function of Y. Then

$$\mathbb{E}|X - Y| = \int_{-\infty}^{\infty} F(x)(1 - G(x))\mathrm{d}x + \int_{-\infty}^{\infty} G(x)(1 - F(x))\mathrm{d}x.$$

Proof. We use

$$|X - Y| = \int_{-\infty}^{\infty} \left[\mathbb{1}_{\{X \leq u < Y\}}(u) + \mathbb{1}_{\{Y \leq u < X\}}(u)\right] \mathrm{d}u$$

and apply Fubini's theorem. This leads to

$$\begin{aligned}
\mathbb{E}|X-Y| &= \mathbb{E}\left(\int_{-\infty}^{\infty}\left[\mathbb{1}_{\{X\leq u<Y\}}(u)+\mathbb{1}_{\{Y\leq u<X\}}(u)\right]\mathrm{d}u\right)\\
&= \int_{-\infty}^{\infty}\mathbb{E}\left(\mathbb{1}_{\{X\leq u<Y\}}(u)\right)\mathrm{d}u+\int_{-\infty}^{\infty}\mathbb{E}\left(\mathbb{1}_{\{Y\leq u<X\}}(u)\right)\mathrm{d}u\\
&= \int_{-\infty}^{\infty}\mathbb{P}(X\leq u<Y)\mathrm{d}u+\int_{-\infty}^{\infty}\mathbb{P}(Y\leq u<X)\mathrm{d}u\\
&= \int_{-\infty}^{\infty}\mathbb{P}(X\leq u)\cdot\mathbb{P}(Y>u)\mathrm{d}u+\int_{-\infty}^{\infty}\mathbb{P}(Y\leq u)\cdot\mathbb{P}(X>u)\mathrm{d}u\\
&= \int_{-\infty}^{\infty}F(u)\left(1-G(u)\right)\mathrm{d}u+\int_{-\infty}^{\infty}G(u)\left(1-F(u)\right)\mathrm{d}u.
\end{aligned}$$

□

Theorem 3.3.4. The CRPS can be stated equivalently as

$$\mathrm{CRPS}(F,y)=\mathbb{E}|X-y|-\frac{1}{2}\mathbb{E}|X-\tilde{X}|, \tag{3.5}$$

where X und X' are independent copies of a random variable with distribution F.

Proof. This statement follows directly form Lemma 3.3.3. □

In the following, we will show that the CRPS is a strictly proper scoring rule relative to the class $\mathcal{P}$ of probability measures on $\mathbb{R}$. For Lemma 3.3.5 we refer to Baringhaus and Franz (2004)

Lemma 3.3.5. Let $X,\tilde{X},Y,\tilde{Y}$ be independent real random variables with finite expectations. Let $X,\tilde{X}$ be identically distributed with distribution function F, and let $Y,\tilde{Y}$ be identically distributed with distribution function G. Then

$$\mathbb{E}|X-Y|-\frac{1}{2}\mathbb{E}|X-\tilde{X}|-\frac{1}{2}\mathbb{E}|Y-\tilde{Y}|\geq 0, \tag{3.6}$$

where equality holds if and only if $F=G$.

Proof. Inequality (3.6) follows from the fact that due to Lemma 3.3.3 the identity

$$\mathbb{E}|X-Y|-\frac{1}{2}\mathbb{E}|X-\tilde{X}|-\frac{1}{2}\mathbb{E}|Y-\tilde{X}|=\int_{-\infty}^{\infty}(F(x)-G(x))^2\mathrm{d}x$$

is true. Equality holds if and only if $F=G$ λ-almost everywhere, where λ is the Lebesgue measure on the Borel sets of $\mathbb{R}$. □

Theorem 3.3.6. Let $\mathcal{P}$ be the class of Borel probability measure on $\mathbb{R}$. Then the CRPS is a strictly proper scoring rule with respect to $\mathcal{P}$

Proof. We consider the associated divergence function of the CRPS

$$\begin{aligned} d(F,G) &= \mathrm{CRPS}(F,G) - \mathrm{CRPS}(G,G) \\ &= \mathbb{E}|X-Y| - \frac{1}{2}\mathbb{E}|X-\tilde{X}| - \left(\mathbb{E}|Y-\tilde{Y}| - \frac{1}{2}\mathbb{E}|Y-\tilde{Y}|\right) \\ &= \mathbb{E}|X-Y| - \frac{1}{2}\mathbb{E}|X-\tilde{X}| - \frac{1}{2}\mathbb{E}|Y-\tilde{Y}|, \end{aligned}$$

where X and $\tilde{X}$ are i.i.d copies of a random variable with distribution function F and Y and $\tilde{Y}$ i.i.d copies of a random variable with distribution function G. By applying Lemma 3.6 the assertion follows. □

The two representations of the CRPS already presented are the most frequently used ones. However, note that the CRPS also can be represented quantile-based.

Definition 3.3.7. The *pinball loss* or *quantile loss* at level $\alpha \in [0,1]$ with a predicted α-th quantile q is defined as

$$\Lambda_\alpha(q,x) = (\alpha - \mathbb{1}_{\{x<q\}})(x-q). \tag{3.7}$$

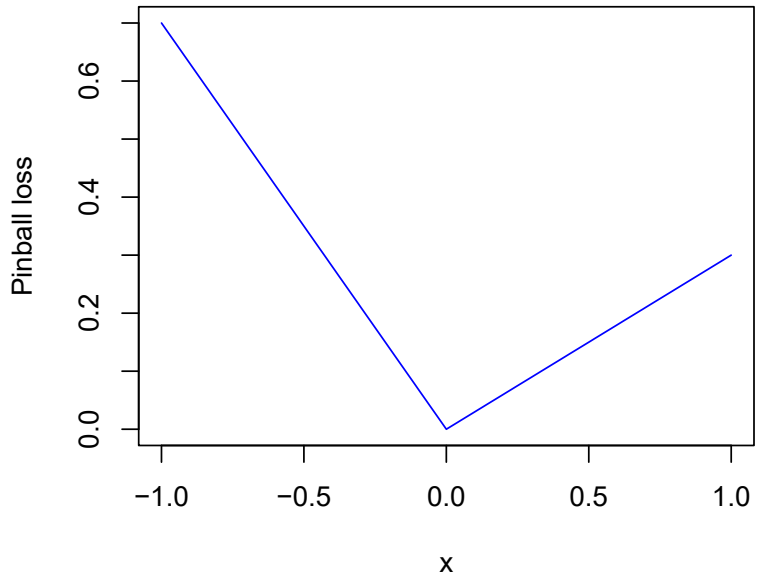


Figure 3.2.: The pinball loss function $\Lambda_0(0.3,x)$

The CRPS can be represented as the pinball loss integrated over all quantile levels $\alpha \in [0,1]$:

$$\mathrm{CRPS}(F,y) = \int_0^1 2\Lambda_\alpha(F^{-1}(\alpha),y)\mathrm{d}\alpha, \tag{3.8}$$

where $F^{-1}(\alpha) = \inf\{x|F(x) \geq \alpha\}$ denotes the α-quantile of F.

The CRPS can also be generalized in order to emphasize certain parts of the forecasting distribution F. Following Definition 3.3.1 we can add a threshold weight function $u : \mathbb{R} \to \mathbb{R}_+$, see Diks, Panchenko, and Van Dijk (2011), so that

$$\mathrm{CRPS}_u(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}_{\{y \leq x\}})^2 u(x) \mathrm{d}x. \tag{3.9}$$

The quantile representation of the CRPS can be generalized as

$$\mathrm{CRPS}_\nu(F, y) = \int_0^1 2\Lambda_\alpha(F^{-1}(\alpha), y)\nu(\alpha)\mathrm{d}\alpha, \tag{3.10}$$

where $\nu : [0, 1] \to \mathbb{R}_+$ is a quantile weight function. Note that Equations (3.9) and (3.10) are in general not equivalent. In general, any non-negative function u and ν can be utilized, provided that Equations (3.9) and (3.10) are convergent.

Emphasis	Threshold weights	Quantile weights
Uniform	$u(x) = 1$	$\nu(\alpha) = 1$
Centre	$u(x) = \phi(x)$	$\nu(\alpha) = \alpha(1 - \alpha)$
Both tails	$u(x) = 1 - \phi(x)/\phi(0)$	$\nu(\alpha) = (2\alpha - 1)^2$
Right tail	$u(x) = \Phi(x)$	$\nu(\alpha) = \alpha^2$
Left tail	$u(x) = 1 - \Phi(x)$	$\nu(\alpha) = (1 - \alpha)^2$

Table 3.1.: Possible weight functions for the CRPS. The weight functions $u : \mathbb{R} \to \mathbb{R}_+$ and $\nu : [0, 1] \to \mathbb{R}_+$ put additional emphasize on certain parts of the forecasting distribution. Forecasts with deviations on these parts are penalized additionally and receive a higher score. Here ϕ and Φ denote the pdf and cdf of the standard normal distribution.

Remark 3.3.8. Note that in this work we are concerned with evaluating multivariate forecasts. With a little adjustment the CRPS can also be applied to multivariate quantities. So let F be the multivariate d-dimensional probabilistic forecast with marginals $F_1, \ldots, F_d$. Then the multivariate extension of the CRPS is definded as

$$\mathrm{CRPS}_{\mathbf{a}}(F, \mathbf{y}) := \sum_{i=1}^{d} a_i \mathrm{CRPS}(F_i, y_i),$$

where $\mathbf{y} = (y_1, \ldots, y_d)$ is the d-dimensional observation and the coefficients $a_1, \ldots, a_d \in \mathbb{R}$ allow us to emphasize or downweight certain coordinates.

3.4. Energy score

The *energy score* is a multivariate generalization of the CRPS and draws on Székely's energy distance, see Székely (2003). This scoring rule is very general and already broadly discussed in literature. The energy score is defined as follows.

Definition 3.4.1. Let $\mathcal{P}_\beta$, $\beta \in (0,2)$, denote the class of the Borel probability measures on $\mathbb{R}^d$ such that $\mathbb{E}\|\mathbf{X}\|^\beta < \infty$. As before, we identify a probabilistic forecast $F \in \mathcal{P}_\beta$ with its (multivariate) cumulative distribution function.
For a probabilistic forecast F and an observation $\mathbf{y} \in \mathbb{R}^d$ the *energy score* is defined as

$$\mathrm{ES}_\beta(F, \mathbf{y}) = \mathbb{E}\|\mathbf{X} - \mathbf{y}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta \tag{3.11}$$

where $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are independent copies of a random vector with distribution F.

Remark 3.4.2. The case $\beta = 1$ seems to be the standard choice in practical application. Obviously, the energy score reduces to the CRPS when $\beta = 1$ and $d = 1$.

In the following, we will show that the energy score is a strictly proper scoring rule relative to the class $\mathcal{P}_\beta$. In literature a variety of proofs of this statement are proposed, see e.g. Gneiting and Raftery (2007).
Here we essentially follow the proof of Baringhaus and Franz (2004). The idea is to reduce the general case $d > 1$ to the special case $d = 1$ using the projection method. Therefore, we assume that $d > 1$ for the rest of this section. Firstly, we have to generalize Lemma 3.3.5.

Lemma 3.4.3. Let $X, \tilde{X}, Y, \tilde{Y}$ be independent real random variables with finite expectations. Let $X, \tilde{X}$ be identically distributed with distribution function F, and let $Y, \tilde{Y}$ be identically distributed with distribution function G. Then for $0 < \beta < 2$

$$\mathbb{E}|X - Y|^\beta - \frac{1}{2}\mathbb{E}_F|X - \tilde{X}|^\beta - \frac{1}{2}\mathbb{E}_F|Y - \tilde{Y}|^\beta \geq 0, \tag{3.12}$$

where equality holds if and only if $F = G$.

Proof. We use the fact that $|X - Y|^\beta$ is a non-negative random variable and substitute $u = t^{\frac{1}{\beta}}$. This yields

$$\begin{aligned}
\mathbb{E}|X - Y|^\beta &= \int_0^\infty \mathbb{P}(|X - Y|^\beta > t)\mathrm{d}t = \int_0^\infty \mathbb{P}(|X - Y| > t^{1/\beta})\mathrm{d}t \\
&= \int_0^\infty \beta u^{\beta-1}\mathbb{P}(|X - Y| > u)\mathrm{d}u \\
&= \int_{\mathbb{R}} \beta|u|^{\beta-1}\left[\mathbb{P}(X < u < Y) + \mathbb{P}(Y < u < X)\right]\mathrm{d}u \\
&= \int_{\mathbb{R}} \beta|u|^{\beta-1}\left[F(u)(1 - G(u)) + G(u)(1 - F(u))\right]\mathrm{d}u.
\end{aligned}$$

Similary, it holds that

$$\begin{aligned}
\mathbb{E}|X - \tilde{X}|^\beta &= \int_{\mathbb{R}} 2\beta|u|^{\beta-1}F(u)(1 - F(u))\mathrm{d}u, \\
\mathbb{E}|Y - \tilde{Y}|^\beta &= \int_{\mathbb{R}} 2\beta|u|^{\beta-1}G(u)(1 - G(u))\mathrm{d}u.
\end{aligned}$$

Therefore, we have

$$2\beta \int_{\mathbb{R}} |u|^{\beta-1}\big(F(u) - G(u)\big)^2 \mathrm{d}u \tag{3.13}$$

$$= 2\beta \int_{\mathbb{R}} |u|^{\beta-1}[F(u)(1-G(u)) + (1-F(u))G(u) - F(u)(1-F(u)) - G(u)(1-G(u))]\mathrm{d}u$$

$$= 2\mathbb{E}|X-Y|^\beta - \mathbb{E}|X-\tilde{X}|^\beta - \mathbb{E}|Y-\tilde{Y}|^\beta. \tag{3.14}$$

The Integral (3.13) is finite if $|\beta - 1| < 1$. Therefore, (3.14) is non-negative und finite for all $0 < \beta < 2$. The equality

$$2\mathbb{E}|X-Y|^\beta - \mathbb{E}|X-\tilde{X}|^\beta - \mathbb{E}|Y-\tilde{Y}|^\beta = 0$$

holds if and only if $F = G$ λ-almost everywhere, where λ is the Lebesgue measure on the Borel sets of $\mathbb{R}$. □

Lemma 3.4.4. For each $\mathbf{x} \in \mathbb{R}^d$ the representation

$$\|\mathbf{x}\| = \gamma_d \cdot \int_{S^{d-1}} |\mathbf{a}'\mathbf{x}| \mathrm{d}\mu(\mathbf{a}) \tag{3.15}$$

holds, where μ is the uniform distribution on $S^{d-1} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$, the surface of the unit sphere in $\mathbb{R}^d$, and

$$\gamma_d = \frac{\sqrt{\pi}(d-1)\Gamma\left(\frac{d-1}{2}\right)}{2\Gamma\left(\frac{d}{2}\right)}. \tag{3.16}$$

Proof. The result is clearly true for $\mathbf{x} = 0$. So let $\mathbf{x} \neq 0$. It is well known, that the uniform distribution on S^{d-1} is invariant with respect to orthogonal transformations, see Bryc (1995). Therefore, we have

$$\int_{S^{d-1}} |\mathbf{a}'\mathbf{x}| \mathrm{d}\mu(a) = \|x\| \int_{S^{d-1}} \left|\mathbf{a}'\frac{\mathbf{x}}{\|\mathbf{x}\|}\right| \mathrm{d}\mu(\mathbf{a}) = \|x\| \int_{S^{d-1}} |\mathbf{a}'\mathbf{e}| \, \mathrm{d}\mu(\mathbf{a}), \tag{3.17}$$

where $\mathbf{e} = (1, 0, \ldots, 0)' \in \mathbb{R}^d$. To calculate the last integral we introduce the standard normal random vector $\xi = (\zeta_1, \ldots, \zeta_d)'$. It applies that $\zeta/\|\zeta\|$ is uniformly distributed on the surface of the unit sphere, see Muller (1959).
Let $\mathbf{X} \sim \mu$. Then it holds for $X_1 := \mathbf{X}'\mathbf{e}$ that

$$X_1 \sim \tau := \frac{1}{\|\zeta\|}\zeta_1 = \frac{\zeta_1}{\sqrt{\zeta_1^2 + \cdots + \zeta_d^2}}.$$

Since $\zeta_1, \ldots, \zeta_d$ are i.i.d standard normally distributed, $\zeta_1^2, \ldots, \zeta_d^2$ are i.i.d. gamma-distributed with shape parameter $\alpha = 1/2$ and rate parameter $\beta = 1/2$. The gamma distribution is closed under convolution. Therefore,

$$\zeta_2^2 + \cdots + \zeta_d^2$$

is gamma-distributed with parameters $\alpha = (d-1)/2$ and $\beta = 1/2$. It follows that

$$\tau^2 = \frac{\zeta_1^2}{\zeta_1^2 + \zeta_2^2 + \cdots + \zeta_d^2}$$

is beta-distributed with parameters $\alpha = 1/2$ and $\beta = (d-1)/2$. The density of $|\tau| = |\mathbf{X}'\mathbf{e}|$ is given by

$$f(x) = \mathbb{1}_{[0,1]}(x) \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} 2\sqrt{1-x^2}^{d-3}. \tag{3.18}$$

To show this we consider the measurable function $g : [0,1] \to [0,\infty)$. We have

$$\begin{aligned}
\mathbb{E}g(|\tau|) &= \mathbb{E}g\left(\sqrt{\tau^2}\right) = \int_0^1 g(\sqrt{x}) \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\cdot\Gamma\left(\frac{d-1}{2}\right)} \frac{(1-x)^{\frac{d-3}{2}}}{\sqrt{x}}\,\mathrm{d}x \\
&= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} \int_0^1 g(\sqrt{x}) \frac{(1-x)^{\frac{d-3}{2}}}{\sqrt{x}}\,\mathrm{d}x \\
&= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} \int_0^1 g(y) \frac{(1-y^2)^{\frac{d-3}{2}}}{y} \cdot 2y\,\mathrm{d}y \\
&= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} \int_0^1 g(y) 2\sqrt{(1-y^2)}^{d-3}\mathrm{d}y
\end{aligned}$$

Thus, $u(y) = \mathbb{1}_{[0,1]}(y) \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} 2\sqrt{(1-y^2)}^{d-3}$ is a density function of $|\tau|$. So we have

$$\begin{aligned}
\mathbb{E}(|\mathbf{X}'\mathbf{e}|) &= \int_{S^{d-1}} |\mathbf{a}'\mathbf{e}|\,\mathrm{d}\mu(\mathbf{a}) \\
&= \int \mathbb{1}_{[0,1]}(y) \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} 2\sqrt{(1-y^2)}^{d-3} y\,\mathrm{d}y \\
&= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} 2 \int_0^1 \sqrt{(1-y^2)}^{d-3} y\,\mathrm{d}y \\
&= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{d-1}{2}\right)} \cdot 2 \cdot \frac{1}{d-1} = \gamma_d^{-1},
\end{aligned}$$

which was to be proved. □

Theorem 3.4.5. Let $\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{Y}, \tilde{\mathbf{Y}}$ be independent d-dimensional random vectors. Let $\mathbf{X}, \tilde{\mathbf{X}}$ be identically distributed with distribution function F and finite expectation $\mathbb{E}\|\mathbf{X}\|^\beta < \infty$, and let $\mathbf{Y}, \tilde{\mathbf{Y}}$ be identically distributed with distribution function G and finite expectation $\mathbb{E}\|\mathbf{Y}\|^\beta < \infty$. Then for $0 < \beta < 2$ the inequality

$$2\mathbb{E}\|\mathbf{X}-\mathbf{Y}\|^\beta - \mathbb{E}\|\mathbf{X}-\tilde{\mathbf{X}}\|^\beta - \mathbb{E}\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^\beta \geq 0 \tag{3.19}$$

is true, and equality holds if and only if $F = G$.

Proof. From Lemma 3.4.3 we have

$$2\mathbb{E}|\mathbf{a}'\mathbf{X} - \mathbf{a}'\mathbf{Y}|^\beta - \mathbb{E}|\mathbf{a}'\mathbf{X} - \mathbf{a}'\tilde{\mathbf{X}}|^\beta - \mathbb{E}|\mathbf{a}'\mathbf{Y} - \mathbf{a}'\tilde{\mathbf{Y}}|^\beta \geq 0 \tag{3.20}$$

for each $\mathbf{a} \in S^{d-1}$. Integrating with respect to the uniform distribution μ on S^{d-1} we obtain (3.19) from Lemma 3.4.4. Equality in 3.19 holds if and only if for μ-almost every $\mathbf{a} \in S^{d-1}$ the distributions of $\mathbf{a}'\mathbf{X}$ and $\mathbf{a}'\mathbf{Y}$ coincide. For each $t \in \mathbb{R}$ the characteristic functions

$$\varphi_{\mathbf{a}'\mathbf{X}} := \mathbb{E}\exp(it(\mathbf{a}'\mathbf{X})), \quad \mathbf{a} \in S^{d-1},$$

and

$$\varphi_{\mathbf{a}'\mathbf{Y}} := \mathbb{E}\exp(it(\mathbf{a}'\mathbf{Y})), \quad \mathbf{a} \in S^{d-1},$$

are continuous. Thus, equality in (3.19) holds if and only if $\mathbf{X}$ and $\mathbf{Y}$ have the same Fourier transform or equivalently the same distribution. □

Corollary 3.4.6. It follows directly from Theorem 3.19, that the energy score ES_β is a strictly proper scoring rule for $\beta \in (0, 2)$ relative to the class $\mathcal{P}_\beta$ of Borel probability measures on $\mathbb{R}^d$ such that $\mathbb{E}||\mathbf{X}||^\beta < \infty$.

Lemma 3.4.7. In the limiting case $\beta = 2$ the energy score reduces to the squared error.

$$\mathrm{ES}_2(F, \mathbf{y}) = \|\mu_F - \mathbf{y}\|^2,$$

where μ_F denotes the mean vector of the distribution F.

Proof. This statement follows from a straightforward calculation:

$$\begin{aligned}
\mathrm{ES}_2(F, \mathbf{y}) &= \mathbb{E}\|\mathbf{X} - \mathbf{y}\|^2 - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^2 \\
&= \sum_{i=1}^{d} \mathbb{E}(X_i - y_i)^2 - \frac{1}{2}\sum_{i=1}^{d} \mathbb{E}(X_i - \tilde{X}_i)^2 \\
&= \sum_{i=1}^{d} \left(\mathbb{E}X_i^2 - 2y_i\mathbb{E}X_i + y_i^2\right) - \frac{1}{2}\sum_{i=1}^{d} \left(2\mathbb{E}X_i^2 - 2(\mathbb{E}X_i)^2\right) \\
&= \sum_{i=1}^{d} (\mathbb{E}X_i)^2 - 2y_i\mathbb{E}X_i + y_i^2 \\
&= \sum_{i=1}^{d} (\mathbb{E}X_i - y_i)^2 = \|\mu_F - \mathbf{y}\tilde{}|^2.
\end{aligned}$$

□

The associated divergence function is given by

$$\begin{aligned} d_{\mathrm{ES}_2}(F,G) &= \mathrm{ES}_2(F,G) - \mathrm{ES}_2(F,F) \\ &= \mathbb{E}\|\mathbf{X}-\mathbf{Y}\|^2 - \frac{1}{2}\mathbb{E}\|\mathbf{X}-\tilde{\mathbf{X}}\|^2 - \frac{1}{2}\mathbb{E}\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^2 \\ &= \sum_{i=1}^{d}\mathbb{E}(X_i-Y_i)^2 - \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}(X_i-\tilde{X}_i)^2 - \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}(Y_i-\tilde{Y}_i)^2 \\ &= \sum_{i=1}^{d}\left(\mathbb{E}X_i^2 - 2\mathbb{E}Y_i\mathbb{E}X_i + \mathbb{E}Y_i^2\right) - \frac{1}{2}\sum_{i=1}^{d}\left(2\mathbb{E}X_i^2 - 2(\mathbb{E}X_i)^2\right) \\ &\quad -\frac{1}{2}\sum_{i=1}^{d}\left(2\mathbb{E}Y_i^2 - 2(\mathbb{E}Y_i)^2\right) \\ &= \sum_{i=1}^{d}(\mathbb{E}X_i - \mathbb{E}Y_i)^2 = \|\mathbb{E}\mathbf{X} - \mathbb{E}\mathbf{Y}\|^2, \end{aligned}$$

where $\mathbf{X}, \tilde{\mathbf{X}} \sim F$ and $\mathbf{Y}, \tilde{\mathbf{Y}} \sim G$.

Remark 3.4.8. Obviously the scoring rule ES_2 is regular and proper as $d_{\mathrm{ES}_2}(F,G) = 0$ if $F = G$, that is

$$\mathrm{ES}_2(F,F) \leq \mathrm{ES}_2(F,G)$$

for all $F, G \in \mathcal{P}_2$. Note that ES_2 is not strictly proper since we have $d_{\mathrm{ES}_2}(F,G) = 0$ whenever $\mathbb{E}\mathbf{X} = \mathbb{E}\mathbf{Y}$, i.e. this score only depends on the first moment of the distribution.

3.5. Variogram score

The *variogram score* was introduced by Scheuerer and Hamill (2015) and is designed particularly to be sensitive to misspecified correlations between the different components of the forecast. The authors also hypothesize it is readily usable for ensemble forecast diagnosis. The variogram score is motivated by the concept of a variogam, which is also referred to as structure function.

Definition 3.5.1. Let $\mathbf{Y}$ be a d-dimensional random variable. The *variogram of order* $p > 0$ is defined as

$$\gamma_p(i,j) := \frac{1}{2}\mathbb{E}|Y_i - Y_j|^p, \quad i,j, \leq d. \tag{3.21}$$

The variogram is a popular tool in geostatistics and considers the pairwise differences of the components of the multivariate variable of interest $\mathbf{Y}$.

Remark 3.5.2. Let the order of the variogram be $p = 2$. Denoting $\mu_i := \mathbb{E}(Y_i)$, $\sigma_i^2 := \mathrm{var}(Y_i)$ and $\rho_{ij} := \mathrm{corr}(Y_i, Y_j)$ we have

$$\mathbb{E}|Y_i - Y_j|^2 = (\mu_i - \mu_j)^2 + \left(\sigma_i^2 - 2\sigma_i\sigma_j\rho_{ij} + \sigma_j^2\right). \tag{3.22}$$

This shows that the variogram γ_2 not only depends on the first two moments of the individual components, but also on their correlations, see Scheuerer and Hamill (2015).

Definition 3.5.3. The special cases $p = 0.5$ and $p = 1$ are referred to as *rodogram* and *madogram*. Variograms of order p can be defined for any multivariate quantity of interest for which the p-th absolute moment exists, see Scheuerer and Hamill (2015).

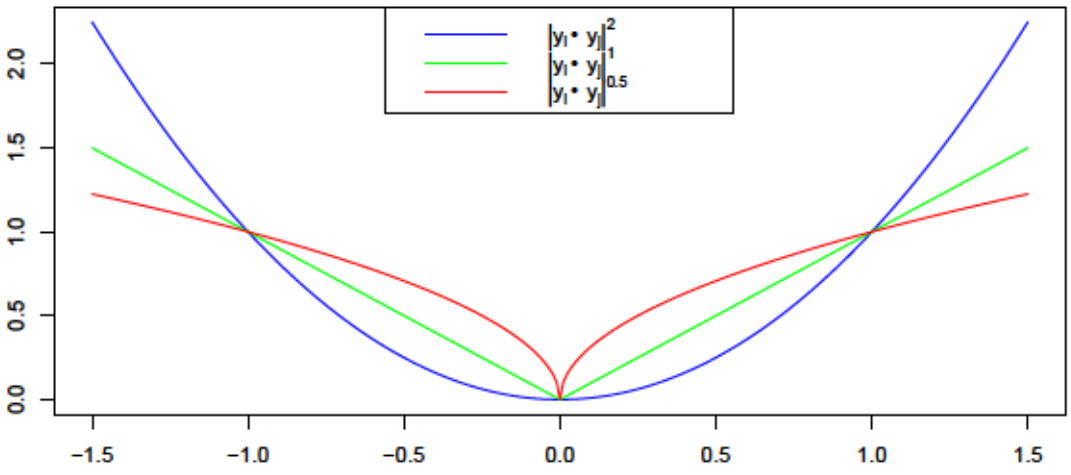


Figure 3.3.: Variogram observations of different orders. The figure shows the effect of the variogram order p on the observed absolute difference $|y_i - y_j|$. Slight deviations of $|y_i - y_j|$ affect the variogram $|y_i - y_j|^p$ differently depending on the order p.

Remark 3.5.4. Note that for $p \neq 2$ and non-Gaussian distributions the variogram can usually not be expressed as a simple function of the means, variances and correlations of Y_i and Y_j, but it still depends on those quantities. Therefore, variograms of order p are potentially useful for comparing the multivariate dependence structure of forecasts and observations.

Utilizing the variograms $\gamma_p(i, j)$ in the context of scoring rules result in a performance measure that is sensitive to different types of miscalibration of multivariate forecasts.

Definition 3.5.5. For a given d-dimensional observation vector $\mathbf{y} = (y_1, \ldots, y_d)$ and a multivariate predictive distribution F the *variogram score of order* $p > 0$ is defined as

$$\mathrm{VS}_p(F, \mathbf{y}) = \sum_{i,j=1}^{d} \omega_{i,j} \big(|y_i - y_j|^p - \mathbb{E}_F |X_i - X_j|^p \big)^2, \tag{3.23}$$

where X_i and X_j are the i-th and j-th component of a random vector $\mathbf{X} \sim F$. The variogram score depends on the choice of the order $p > 0$ and the non-negatives weights $\omega_{i,j}$.

The variogram score VS_p measures the dissimilarity between approximations of the variograms of order p of observations and forecasts over all pairs of components of the variable of interest, see Scheuerer and Hamill (2015).
In applications standard choices for the order of the variograms are $p = 0.5$ or $p = 1$. Also certain pairs of squared variogram differences can be emphasized or downweighted by the choice of weights ω_{ij}. Common choices for the weights are $\omega_{ij} = 1$ or inverse distance weights. Especially in situations where there is some notion of the distance between the i-th and j-th component the latter choice seems reasonable, as correlations at short distances are typically stronger than those at larger distances.

Theorem 3.5.6. The variogram score is a proper scoring rule relative to the class of probability distributions with finite $(2p)$-th moments of all components.

Proof. We consider a pair (i, j). For each pair the mean of the random variable $Z := |X_i - X_j|^p$ minimizes the expected squared deviation of Z for any fixed number $\mathbf{a} \in \mathbb{R}$, i.e.

$$\mathbb{E}(Z - \mathbb{E}(Z))^2 \leq \mathbb{E}(Z - a)^2. \tag{3.24}$$

This means inequality in (3.2) holds separately for any pair (i, j), but then it also holds for the weighted sum over all pairs, for any choice of the non-negative weights. □

Remark 3.5.7. Note that the variogram score is not strictly proper. For instance, large-scale random errors that are the same for every component cancel out when differences are considered. Also a bias that is the same for all components will not be recognized which can be seen directly from the definition of the variogram score, see Scheuerer and Hamill (2015). So particularly a shift of the forecasting distribution in comparison to the true distribution will not be recognized.

3.6. Dawid-Sebastiani score

The *Dawid-Sebastiani score* is motivated by the *logarithmic score* (or *log-score*). The log-score was proposed by Good (1992) and is defined as follows. Firstly let μ be a σ-finite measure on the sample space $(\Omega, \mathcal{A})$. A probabilistic forecast $F \in \mathcal{L}_1$ is then identified with its μ-density f, where $\mathcal{L}_1$ denotes the class of probability measures dominated by μ.

Definition 3.6.1. The *logarithmic score* for the forecasting density f and observation $y \in \Omega$ is defined as

$$\mathrm{LogS}(f, y) := \log(f(y)). \tag{3.25}$$

Theorem 3.6.2. The logarithmic score is strictly proper relative to the class $\mathcal{L}_1$ of probability measures dominated by μ.

Proof. The associated divergence function becomes the classical Kullback-Leibler divergence, see Kullback and Leibler (1951), so it follows directly that the logarithmic score is strictly proper relative to the class $\mathcal{L}_1$. □

Remark 3.6.3. Note that to apply this scoring rule we have to have a probabilistic forecast available taking the form of a predictive density.

Based on the logarithmic score Dawid and Sebastiani proposed a scoring rule that applies to the Borel probability measures on $\mathbb{R}^d$ and that depends on the predictive distribution F only through its mean μ and its covariance matrix Σ.

Definition 3.6.4. Let F be a Borel probability measure on $\mathbb{R}^d$ with mean μ and covariance matrix Σ. The *Dawid-Sebastiani score* for the forecasting distribution F and observation $\mathbf{y} \in \mathbb{R}^d$ is defined as

$$\mathrm{DSS}(F, \mathbf{y}) := \log\left(\det\left(\Sigma\right)\right) + \left(\mathbf{y} - \mu\right)' \Sigma^{-1} \left(\mathbf{y} - \mu\right). \tag{3.26}$$

Remark 3.6.5. The first term penalizes dispersion (i.e. lack of sharpness) of the ensemble. The second term penalizes lack of correspondence between the observed and ensemble-mean vectors through the Mahalanobis distance, see Mardia, Kent, and Bibby (1979).

Remark 3.6.6. Note that the Dawid-Sebastiani score can take values on the real line, that is $-\infty < \mathrm{DSS} < \infty$.

Theorem 3.6.7. The Dawid-Sebastiani score is strictly proper relative to the convex class of Gaussian measures.

Proof. Note that Equation 3.26 is a linear transformation of the log-likelihood function for the multivariate Gaussian distribution, and so is related to the logarithmic score. Therefore, it is strictly proper for Gaussian variables. □

Remark 3.6.8. The Dawid-Sebastiani score can also be computed for forecasts and observations where the first two moments of the underlying distributions are finite. In this case it is proper but not strictly proper, see Gneiting and Raftery (2007).

Remark 3.6.9. The divergence function correspoding to the Dawid-Sebastiani score is given by

$$d(F, G) = \mathrm{tr}(\Sigma_F^{-1}\Sigma_G) - \log\left(\det(\Sigma_F^{-1}\Sigma_G)\right) + (\mu_F - \mu_G)'\Sigma_F^{-1}(\mu_F - \mu_G) - d,$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

4. Reporting of probabilistic forecasts

After introducing the theoretical concepts of probabilistic forecasts and their evaluation by proper scoring rules we now deal with the reporting of the forecasting distribution in practice.
As already stated above, we assume that the sample space is given by $\Omega_Y = \mathbb{R}^d$. Moreover, let $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ be the d-dimensional random variable of interest with cumulative distribution function G. We observe the realization $\mathbf{y} = (y_1, \ldots, y_d)$ of $\mathbf{Y}$. Further, let $\mathbf{X} = (X_1, \ldots, X_d)$ be the forecast vector for $\mathbf{Y}$ with distribution F.
In practice there are essentially two options of reporting the distribution F that are commonly employed, namely reporting the forecasting distribution F explicitly or reporting the forecast in form of an ensemble, see Gneiting, Stanberry, et al. (2008).

4.1. Explicit representation of the forecasting distribution

This reporting option is quite self explaining. If we utilze this option, we have to report the forecasting distribution F explicitly in form of its CDF, i.e. we report an estimate of the CDF F of the underlying CDF G. This option is very general as it doesn't require any further conditions on the distribution.
It is also possible to report a term that provides equivalent information as the CDF. If the distribution is (absolute) continuous with respect to the Lebesgue measure, it admits a probability density function. Thus, reporting an estimate f of the density g corresponding to the CDF G displays the same information as reporting the CDF in this case. This forecasting technique corresponds to the term density forecast and is commonly utilized in economics and finance, see Clements (2005).
As we are concerned particularly with the interdependence structure of multivariate forecasts we also employ the following reporting option. The forecasting distribution can be described using copulas. According to Sklar's theorem we can decompose every multivariate distribution in its marginals and a copula, which describes the dependency structure, i.e.

$$F(x_1, \ldots, x_d) = C_F(F_1(x_1), \ldots, F_d(x_d))$$

and

$$G(x_1, \ldots, x_d) = C_G(G_1(x_1), \ldots, G_d(x_d)).$$

That is, we can report the forecasting distribution by estimates $F_1, \ldots, F_d$ of the marginals $G_1, \ldots, G_d$ and a copula estimate $C_F(\cdot)$ of $C_G(\cdot)$.
Another option is to report an estimate of the characteristic function $\psi_{\mathbf{X}}(t) = \mathbb{E}(\exp(i\mathbf{t}'\mathbf{X}))$ of $\psi_Y(t) := \mathbb{E}(\exp(i\mathbf{t}'\mathbf{Y}))$ which uniquely determines the distribution of $\mathbf{Y}$.
Summing up, we report an estimate of the CDF or a term that yields equivalent information. To this end, it has to be noted that from an evaluation point of view we must be able to draw a random sample from the reported distribution or solve some characteristics such as the first two moments of the distribution.
For more sophisticated forecasting models reporting the distribution explicitly can be very challanging or even impossible. However, the reporting option described in the following section can still be applied in these cases. Accordingly, it has become the standard alternative in practical application.

4.2. Ensemble forecast

Basically, an ensemble forecast is a collection of point forecasts for a specific quantity or event. The ensemble forecast has been widely used in weather forecasting and climate prediction, see Palmer (2002). Based on the original application in weather and climate prediction such an ensemble prediction system consists of multiple runs of numerical weather prediction models which differ either in initial conditions or in the model's parameterized analytical expression, see Gneiting and Katzfuss (2014). As the models in this field are generally sophisticated the sample size is small. Typically, between $m = 10$ and $m = 50$ runs are performed since these models require high computing power, see Gneiting and Katzfuss (2014).
In the following, we denote an ensemble forecast by $\mathcal{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)})$, where $\mathbf{X}^{(i)}$ is the outcome of one simulation run.
In practice, there are different approaches in building and interpreting these ensemble models. Ideally, we would consider the ensemble system $\mathcal{X}$ as a collection of equally likely scenarios, drawn from the same distribution. In this case the ensemble forecast represents a Monte-Carlo simulation, particularly if the sample size m is large.
Therefore, one could argue that the distinction between the two described reporting options is somewhat artifical. We can draw from a predictive distribution to obtain a forecast ensemble and conversely estimate a predictive distribution given an ensemble forecast system, particularly if the ensemble size is large.
However, this approach is rarely feasible in practice as ensemble forecasts are subject to biases and dispersion erros, and, thereby call for statistical postprocessing, see Gneiting and Katzfuss (2014). In general, there is no need to assume that the ensemble members are equally likely or even draws from distributions at all. In literature, this methodology is also known as ensemble simulation, path simulation, trajectory simulation or scenario simulation. Depending on the field the meaning might differ as well.
However, in this work we will treat the ensemble forecast like a Monte-Carlo simulation. Hence, we assume that the members of the ensemble forecast system are an i.i.d. sample, and the sample size m is large. By these assumptions, it holds that the

true underlying distribution function is described well by the simulated sample. This approximation dates back to the multivariate Glivenko-Cantelli theorem. It states that the multivariate empirical CDF is converging almost surely to the drawn distribution.

5. Estimation of the score values

In this section we consider the estimators for the score values in practical application. In this thesis we assume the observation $\mathbf{y}$ of the true distribution G is given and we have the ensemble forecast $\mathcal{X}$ in form of an i.i.d. sample $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)}$ available, that is $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)}$ are independent draws from the forecasting distribution F. Following Ziel and Berk (2019), we formulate estimators for the different scoring rules.

5.1. CRPS

Basically, there are two options to estimate the CRPS which correspond to the different representations (3.4) and (3.5). As the CRPS is a univariate scoring rule, we can only compute the score for each marginal of the forecast and the observation.
By utilizing (3.4) one might be able to solve the corresponding integral in closed form if the predictive distribution is reported explicitly.
If the forecast is reported in form of an i.i.d. sample $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)}$ the integral can be solved numerically by replacing the distribution function with an empirical distribution function. The resulting estimator for the j-th component is given by

$$\widehat{\mathrm{CRPS}}_j := \int \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{X_j^{(i)} \leq z\}} - \mathbb{1}_{\{y_j \leq z\}} \right)^2 \mathrm{d}z.$$

Alternatively, we can use (3.5) to estimate the two expectations. The resulting estimator is a special case of the estimator of the energy score, see e.g. Taieb et al. (2016), which we discuss in the following subsection.
Given the scores of each marginal we can compute the corresponding marginal score by

$$\widehat{\mathrm{CRPS}}_{\mathbf{a}} := \sum_{j=1}^{d} a_j \cdot \widehat{\mathrm{CRPS}}_j$$

with constants $a_1, \dots, a_d \in \mathbb{R}$.

5.2. Energy score

If the forecasting distribution is reported explicitly one might be able to provide a closed-form formula for the energy score, for instance, in the case of multivariate Gaussian distributions, see Pinson (2013), and Appendix A.3.

In general, we have to estimate the two terms in (3.11) given the random sample $\mathcal{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)})$ of distribution F.
The calculation of the first part of the energy score is straightforward. We estimate $\mathbb{E}\|\mathbf{X} - \mathbf{y}\|^\beta$ as the sample mean

$$\frac{1}{m} \sum_{j=1}^{m} \|\mathbf{X}^{(j)} - \mathbf{y}\|^\beta.$$

The second part of the energy score $\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta$ has multiple plausible options for estimation as the definition implies that we require the independent copie $\tilde{\mathbf{X}}$ of the forecasting distribution.
By our assumption the members of the sample $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(m)}$ are i.i.d. Therefore, we might use one half of this set as draws from $\mathbf{X}$ and the other half as draws from $\tilde{\mathbf{X}}$. So the resulting estimator is given by

$$\frac{1}{\lfloor 0.5m \rfloor} \sum_{j=1}^{\lfloor 0.5m \rfloor} \|\mathbf{X}^{(j)} - \mathbf{X}^{(\lfloor 0.5m \rfloor + j)}\|^\beta$$

Note that the sum only contains $\lfloor m/2 \rfloor$ summands but they have nice statistical properties as they are i.i.d.
Another reasonable estimator is given by

$$\frac{1}{m \cdot (m-1)} \sum_{j=1}^{m} \sum_{k=1}^{m} \|\mathbf{X}^{(j)} - \mathbf{X}^{(k)}\|^\beta.$$

This estimator has the advantage of using a larger amount of summands for approximating the sum which should increase the precision of the estimator in general. Note that the elements of the sum become pairwise dependent, so this improvement is weaker as if the summands were independent.
The higher accuracy of the estimator goes hand in hand with a higher computational demand. The amount of summands increases quadratically in m. Altogether, estimators for the energy score are given by

$$\widehat{\mathrm{ES}}_\beta^{iid} = \frac{1}{m} \sum_{j=1}^{m} \|\mathbf{X}^{(j)} - \mathbf{y}\|^\beta - \frac{1}{2} \cdot \frac{1}{\lfloor 0.5m \rfloor} \sum_{j=1}^{\lfloor 0.5m \rfloor} \|\mathbf{X}^{(j)} - \mathbf{X}^{(\lfloor 0.5m \rfloor + j)}\|^\beta$$

and

$$\widehat{\mathrm{ES}}_\beta^{band} = \frac{1}{m} \sum_{j=1}^{m} \|\mathbf{X}^{(j)} - \mathbf{y}\|^\beta - \frac{1}{2} \cdot \frac{1}{m \cdot (m-1)} \sum_{j=1}^{m} \sum_{k=1}^{m} \|\mathbf{X}^{(j)} - \mathbf{X}^{(k)}\|^\beta.$$

5.3. Variogram score

The variogram score can be computed analogously to the energy score. If we apply Reporting Option 1, we might be able to find closed-form solutions for the score value

in some special case, like the case of multivariate normality.
In the standard case we can estimate the expectation (3.23) by the sample mean. Thus, for a given weight matrix $W = (\omega_{i,j})_{i,j=1,\dots,d} \in \mathbb{R}_+^{d\times d}$ with non-negative weights and order $p > 0$ we have

$$\widehat{VS}_{W,p} = \sum_{i,j=1}^{d} \omega_{i,j} \left(|y_i - y_j|^p - \frac{1}{m} \sum_{k=1}^{m} |X_i^{(k)} - X_j^{(k)}|^p \right)^2 .$$

Note that because of the symmetries $|y_i - y_j| = |y_j - y_i|$ and $|X_i^{(k)} - X_j^{(k)}| = |X_j^{(k)} - X_i^{(k)}|$ computation costs can be halved in the implementation due to

$$\widehat{VS}_{W,p} = 2 \cdot \sum_{i=1}^{d} \sum_{j=i+1}^{d} \omega_{i,j} \left(|y_i - y_j|^p - \frac{1}{m} \sum_{k=1}^{m} |X_i^{(k)} - X_j^{(k)}|^p \right)^2 . \tag{5.1}$$

5.4. Dawid-Sebastiani score

The Dawid-Sebastiani score only depends on the first two moments of the forecasting distribution. If Reporting Option 1 is used, we are usually able to compute these moments explicitly.
In the case of Reporting Option 2 we can estimate the first two moments by their sample counterparts

$$\begin{aligned} \hat{\mu}_{\mathbf{X}} &= \frac{1}{m} \sum_{k=1}^{m} \mathbf{X}^{(k)} \\ \hat{\Sigma}_{\mathbf{X}} &= \frac{1}{m-1} \sum_{k=1}^{m} \left(\mathbf{X}^{(k)} - \hat{\mu}_{\mathbf{X}} \right) \left(\mathbf{X}^{(k)} - \hat{\mu}_{\mathbf{X}} \right)' . \end{aligned}$$

Note that the sample mean and the sample covariance matrix are unbiased estimators. So the estimator for the Dawid-Sebastiani score is given by

$$\widehat{\mathrm{DSS}} = \log \left(\det \left(\hat{\Sigma}_{\mathbf{X}} \right) \right) + \left(\mathbf{y} - \hat{\mu}_{\mathbf{X}} \right)' \hat{\Sigma}_{\mathbf{X}}^{-1} \left(\mathbf{y} - \hat{\mu}_{\mathbf{X}} \right) .$$

6. Evaluation of predictive performance

In this chapter we discuss different measurements of the predictive performance of a probabilistic forecast.
These measures not only allow to evaluate the predictive performance of different forecasts but also enable us to assess the discrimination ability, see Chapter 7, of different scoring rules.
In this thesis we consider the following setting: Let G be the true underlying distribution which we aim to forecast by the forecasting distribution F. We assume that we have the observations $\mathbf{y}_1, \dots, \mathbf{y}_N$ as random draws from the underlying distribution G available.
Furthermore, we have the sample $\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(m)}$ for $i = 1, \dots, N$ available, where each $\mathbf{X}_i^{(k)}$ is a random draw from the forecasting distribution F. In the following, we denote the estimated score corresponding to the forecasting distribution F, and the observation $\mathbf{y}_i$ of G as

$$\widehat{\mathrm{Sc}}(F, \mathbf{y}_i),$$

where we utilize the random sample $\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(m)}$ drawn from distribution F, and the estimator $\widehat{\mathrm{Sc}}$ corresponding to the scoring rules Sc. The estimators of the scoring rules considered in this thesis are discribed in Section 5. As an abbreviation we write $\mathrm{Sc}_i^F := \widehat{\mathrm{Sc}}(F, \mathbf{y}_i)$. The sample mean

$$\overline{\mathrm{Sc}}^F = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Sc}_i^F$$

will be the most relevant criterion when comparing the predictive performance of two forecasting models.

6.1. Relative change in score

The *relative change in score value* is employed in Pinson and Tastu (2013). It consideres the relative change in score value of a forecasting distribution F with respect to the perfect forecast, where the forecaster reports the true underlying distribution correctly. So let the underlying distribution be given by G, then the relative change in score is

defined as

$$\Delta\mathrm{Sc}(F,G) = \frac{\overline{\mathrm{Sc}}^{F} - \overline{\mathrm{Sc}^{G}}}{\overline{\mathrm{Sc}^{G}}},$$

when the forecasting distribution is given by F. The score value $\overline{\mathrm{Sc}^{G}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Sc}_i^G$ comes directly from the inherent uncertainty of the random variable which we aim to forecast, see Pinson and Tastu (2013).
Note that in this definition the relative change in score is formulated in terms of the empirical sample means of score values.
Alternatively, the relative change in score can also be defined in terms of the expected score values, i.e.

$$\mathrm{RelChange}(F,G) = \frac{\mathrm{Sc}(F,G) - \mathrm{Sc}(G,G)}{\mathrm{Sc}(G,G)},$$

which we will denote as the expected relative change in score.

6.2. Generalized discrimination heuristic

The relative change in score value can be adjusted to measure the discrimination ability of scoring rules over multiple misspecified forecasting models, see Alexander et al. (2022).
This *generalized discrimination heuristic* is defined as

$$\mathrm{GDH}^{\mathrm{Sc}} = \frac{1}{k}\sum_{i=1}^{k}\frac{\overline{\mathrm{Sc}^{F^{(i)}}}}{\overline{\mathrm{Sc}^{G}}},$$

where $F^{(1)},\ldots,F^{(k)}$ are different forecasting distributions. A large value for the generalized discrimination heuristic may indicate that the ranking of the scores is reliable and robust in the simulation of sample size N.

6.3. Error rate comparison

In Alexander et al. (2022) another metric for assessing the discrimination ability of a scoring rule is proposed. The *error rate* is the probability of the event that the score corresponding to a miscalibrated forecast is lower than the score corresponding to the true model, i.e. the error rate is the probability that the score difference

$$\mathrm{Sc}_i^F - \mathrm{Sc}_i^G$$

is lower than zero for $i = 1,\ldots,N$.
Note that the error rate is a binary metric and does not take the magnitude into account by which the scores of misspecified models are smaller than that of the true model. By averaging over a sample of scores the error rate decreases, until it reaches

zero in case of strictly proper scoring rules as the sample size N increases.
Furthermore, the distribution of the absolute differences $\mathrm{Sc}_1^F - \mathrm{Sc}_1^G, \ldots, \mathrm{Sc}_N^F - \mathrm{Sc}_N^G$ can be studied.

6.4. Test for significance: the Diebold-Mariano test

The most important criterion for comparing probabilistic forecasts will be the *Diebold-Mariano test*. To assess the quality of different probabilistic forecasts or the discrimination ability of a scoring rule the comparison of the sample means or the relative change in score is incomplete as no consideration is given to the statistical significance of the result. In any particular realization one forecast has to perform better, but one wants to know if the difference in score value is statistically significant. Therfore, we want to know whether a lower score value 'in sample' was good luck, or truely an indicator of a difference 'in population', see Diebold (2015).
Hence, the need for a formal test for comparing predictive accuracy is obvious. Thus, we want to give an overview of a test introduced by Diebold and Mariano that was designed for point forecast evaluation, see Diebold and Mariano (2002). The test design is very general and allows for several generalizations. In Weron and Ziel (2019) and Uniejewski, Weron, and Ziel (2017) the Diebold-Mariano test is applied in multivariate settings and in Möller, Lenkoski, and Thorarinsdottir (2013) the test is utilized in an application for the energy score.
In the following, we describe the Diebold-Mariano test exactly for our setting. The true distribution is given by G, and we have the two competing forecasts $F^{(1)}$ and $F^{(2)}$ available. From an evaluation point of view we have the observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$ as random draws from distribution G and the i.i.d. sample $\mathbf{X}_i^{(1)}, \ldots, \mathbf{X}_i^{(m)}$ for $i = 1, \ldots N$ as random draws from the distribution $F^{(1)}$ as well as the i.i.d. sample $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}_i^{(m)}$ for $i = 1, \ldots N$ as random draws from the distribution $F^{(2)}$ available.
The DM-test checks if the corresponding sample mean $\overline{\mathrm{Sc}}^{F^{(1)}}$ is significantly different from $\overline{\mathrm{Sc}}^{F^{(2)}}$. Therefore, score differences are required. Given the losses $\mathrm{Sc}_1^{F^{(1)}}, \ldots, \mathrm{Sc}_N^{F^{(1)}}$ and $\mathrm{Sc}_1^{F^{(2)}}, \ldots, \mathrm{Sc}_N^{F^{(2)}}$ we define the loss differential between the two forecasts as

$$d_j^{F^{(1)}F^{(2)}} = \mathrm{Sc}_j^{F^{(1)}} - \mathrm{Sc}_j^{F^{(2)}}, \quad \text{for } j = 1, \ldots, N.$$

In this way, we can state the null hypothesis of the Diebold-Mariano test as

$$H_0: \quad \mathbb{E}(d_j^{F^{(1)}F^{(2)}}) = 0 \quad \text{for all } j = 1, \ldots, N$$

versus the alternative hypothesis

$$H_1: \quad \mathbb{E}(d_j^{F^{(1)}F^{(2)}}) \neq 0.$$

The null hypothesis states that the competing forecasts have the same accuracy, whereas the alternative hypothesis says that the forecasts have different levels of accuracy.

The asymptotic test is constructed as follows. For the two forecasts we have the sample loss differential $\left(d_j^{F^{(1)}F^{(2)}}\right)_{j=1}^{N}$.

Note that in the setting of this thesis the loss differential series $\left(d_j^{F^{(1)}F^{(2)}}\right)_{j=1}^{N}$ is independent. So under the null hypothesis H_0 it holds that

$$\frac{\overline{d}-\mu}{\sqrt{\frac{\sigma^2}{N}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

- $\mu = \mathbb{E}(d_j^{F^{(1)}F^{(2)}})$,
- $\overline{d} = \frac{1}{N}\sum_{j=1}^{N} d_j^{F^{(1)}F^{(2)}}$ is the sample mean of the loss differential, and
- σ^2 is the variance of the loss differential.

This result follows directly by the central limit theorem. Resulting from this the Diebold-Mariano test utilizes the test statistic

$$\mathrm{DM} = \frac{\overline{d}}{\sqrt{\frac{\hat{\sigma}^2}{N}}},$$

where $\hat{\sigma}^2$ is sample covariance of the loss differential. The null hypothesis of the Diebold-Mariano test is rejected at a significance level α if $|\mathrm{DM}| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ - quantile of the standard normal distribution.

7. Discrimination ability

In this thesis we are mainly concerned with the discrimination ability of the different scoring rules. This means we want to know how well the scoring rules can distinguish between different forecasting distributions.
The concept of discrimination ability is inspired by the work of Murphy (1993). Note that the concept of discrimination ability is a property of the scoring rule only, not of the forecasts themselves.
In general, the discrimination ability can not be assessed by the definition of the scoring rule. We can only prove that a scoring rule is strictly proper.
The most important measurments to assess the discrimination ability of a scoring rule is the Diebold-Mariano test as all other measures do not consider the statistical significance of the results.

7.1. Introduction

Propriety is a property of scoring rules involving a predictive distribution and the real distribution of the underlying process. In practice, the real distribution is generally unknown, and one is left with comparing alternative forecasting distributions, say $F^{(1)}$ and $F^{(2)}$, generated by two rival forecasters.
Propriety does not ensure that a difference in quality of the two forecasts $F^{(1)}$ and $F^{(2)}$ would yield a difference between the resulting score values $\mathrm{Sc}(F^{(1)}, \mathbf{y})$ and $\mathrm{Sc}(F^{(2)}, \mathbf{y})$ for any observation $\mathbf{y} \in \mathbb{R}^d$ drawn from the true distribution G.
Consequently, Pinson and Tastu (2013) refer to discrimination ability as the property of the scoring rule Sc such that

$$F^{(1)} \succ F^{(2)} \Longleftrightarrow \mathrm{Sc}(F^{(1)}, \mathbf{y}) < \mathrm{Sc}(F^{(2)}, \mathbf{y}) \tag{7.1}$$

for any observation $\mathbf{y}$ drawn from G. In the above $F^{(1)} \succ F^{(2)}$ means that $F^{(1)}$ is a forecast of higher quality than $F^{(2)}$. A scoring rule is said to have a higher discrimination ability if differences in quality between predictive distributions reflect in significant differences in score values. On the opposite, a scoring rule is said to have no discrimination ability, if the same score values are assigned to predictive distributions of different quality.
Note that proper scoring rules as the variogram score may not need to have any discrimination ability at all, since it is possible that they assign the same score values to all predictive distributions F, as well as to the true underlying distribution G from which the observation $\mathbf{y}$ was drawn. The situation is different for strictly proper scoring

rules. They should at least discriminate between forecasting distributions that are in the neighbourhood of G. However, for probabilistic forecasts significantly different from the underlying distribution it is not safe to claim that the scoring rule discriminates among predictive distributions.

7.2. Preliminaries on discrimination ability of different scoring rules

Although the energy score is the most prominent multivariate strictly proper scoring rule, it is frequently criticized in literature due to the apparently poor discrimination ability with respect to the interdependence structure of probabilistic forecasts, see e.g. Pinson and Tastu (2013), and Scheuerer and Hamill (2015).
In the study of Pinson and Tastu (2013) the discrimination ability of the energy score is evaluated in a simulation study while considering the bivariate normal distribution. In their simulation study the authors utilized the relative change in score value to assess the discrimination ability of the energy score. They conclude that the energy score can not separate differences in the dependency structure well.
In this working paper also an upper bound of the relative change in score of the energy score in the multivariate Gaussian case is calculated. To this end, they compute the relative change in score for the naive forecast which totally ignores the interdependence between the components, while for the true distributions the components are perfectly dependent. In this case they are able to find a closed-form solution for the relative change in score.
They found that this upper bound is increasing in dimension size d, but it reaches an asymptote of less than 15 % for higher dimensions, see appendix A.3 with σ and β both set to 1.
So overall, this made reseachers and practitioners skeptical in using the energy score.
In the following, we will discuss this simulation study in detail and draw some different conclusions about it.

Following the conclusions of Pinson and Tastu (2013) and the resulting skepticism about the discrimination ability of the energy score, new scoring rules have been proposed. The most prominent example is the variogram score which was introduced to overcome the reported problems of the energy score, see Scheuerer and Hamill (2015). It is evident by its definition that this scoring rule is specifically designed to be sensitive with respect to the interdependence structure of the forecasting distribution. However, the variogram score has one major drawback, namely it is not strictly proper. Thus, it can not identify the true underlying model.
It is also problematic that large-scale random errors that are the same for all components cancel out when differences are considered; likewise, a bias that is the same for all components will go undetected. Furthermore, the variogram score of order p is not able to distuingish among distributions which have the same p-th moments but differ

in higher moments.

Another reported plausible candidate is the multivariate log-score. This scoring rule requires a multivariate density forecast which is not the case in many applications. As discussed in A. Jordan, Krüger, and Lerch (2017), this density can be approximated. Let us remark that these approximation methods suffer efficiency in higher dimension and the results depend crucially on the chosen approximation methods.

Furthermore, we have the Dawid-Sebastiani score, see e.g. Gneiting and Raftery (2007). This scoring rule evaluates the mean and covariance matrix of the distributions and corresponds to the log-score in the multivariate Gaussian setting. Moreover, a characterization only by the first two moments is not sufficient in many applications. In general, it holds true that the Dawid-Sebastiani score is not strictly proper.

To address these considerations, Ziel and Berk (2019) introduced a new sort of scoring rule that is sensitive to the interdependence structure of probabilistic forecasts using copula theory.
The idea is to describe a given multivariate distribution by its marginals and the copula for the dependency structure. By extracting the dependency structure the authors intend to construct a more sensitive measurement for evaluating dependencies. Afterwards, by combining a score for the marginals with one for the copula one obtains a proper or even strictly proper scoring rule. However, the authors conclude that these marginal-copula scores do not perform better than the original energy score in terms of discrimination ability based on their simulation studies. This conclusion is not surprising as the dependence structure given by the copula has to be assessed with one of the available multivariate scoring rules. Note that the copula itself is a multivariate distribution.
It also has to be mentioned that this paper is one of the very few that evaluates the discrimination ability of a scoring rule with the Diebold-Mariano test.

Another newly proposed scoring rule is the CRPS-sum introduced by Salinas et al. (2019). Various studies have shown that the energy score in particular, as well as the variogram score and the Dawid-Sebastiani score seem to be not relieable metrics for evaluating multivariate forecasts. Therefore, the CRPS-sum has gained a lot of prominence lately, see Salinas et al. (2019).
While the CRPS-sum has been well-received in the scientific community, the properties of the CRPS-sum are studied merely in Koochali et al. (2022). One major drawback of the CRPS-sum is, that it is not strictly proper. Thus, it is not able to identify the true model. Furthermore, the authors found that the CRPS-sum performs significantly worse than the energy score in their simulation studies, see Salinas et al. (2019). for this reason, and because this scoring rule is not strictly proper we will not consider it any further.

However, when we consider the results of a couple of researches, e.g. Pinson and Tastu (2013) and Scheuerer and Hamill (2015), that mention the weak discrimination ability of the energy score they all refer to the working paper simulation study of Pinson and Tastu (2013). This study only considers the energy score with parameter $\beta = 1$ and utilizes the relative change in score value as a measure to evaluate the discrimination ability of a given scoring rule.
From a statistical point of view a comparison of the sample mean or the relative change of the score values is incomplete as no consideration is given to the statistical significance of the result.
In the following, we want to determine if the difference in the discrimination ability of the scoring rules is statistically significant. In order to study the discrimination ability of the different scoring rules we have to access the score values of forecasting distributions of different quality with the Diebold-Mariano test. If a scoring rule discriminates well
between different forecasting distributions, the Diebold-Mariano test should detect that the forecasting distributions differ significantly.
At this point, we would like to emphasize that the following simulation studies are designed in such a way that we do not want to find the best probabilistic forecast for an unknown distribution. Instead, we want to compare the different scoring rules with respect to their discrimination ability. Therefore, we assume that both the forecasting distributions and the underlying distribution are known.

7.3. Simulation study I: bivariate Gaussian process

As already stated above, Pinson and Tastu (2013) conclude that the energy score nicely discriminates predictive distributions with different mean parameters, but has a poor discrimination ability with respect to the interdependence structure of different forecasts.
In their working paper, the authors considered solely the energy score with parameter $\beta = 1$ and assessed the score values corresponding to the different forecasts with the relative change in score. If the relative change in score value with respect to the true model of a miscalibrated forecast is small, this means that the discrimination ability of the scoring rule is rather poor.
However, note that by simply considering this measurement the statistical significance of the results is not considered. Therefore, we extend the simulation study and assess the score values of the different forecasts utilizing the Diebold-Mariano test with respect to the true distribution.
In this study we not only consider the energy score with parameter $\beta = 1$ because we conjecture that the choice of the parameter β also affects the discrimination ability of the energy score.
Furthermore, we consider the variogram score und the Dawid-Sebastiani score which is designed specifically for Gaussian distributions in order to compare them with the energy score.

For the calculation of the score values an ensemble size of $m = 2^{14}$ is utilized. As a result, we can assume that the forecasting distribution is described well by the ensemble. In contrast, Pinson and Tastu (2013) utilized an ensemble size of $m = 1000$.
Further, this experiment is replicated $N = 2^9$ times, i.e. we use $N = 2^9$ random draws from the underlying distribution $G_{\mathbf{Y}}$.
Analogous to the study of Pinson and Tastu (2013), let the distribution generating the real process be given by a bivariate Gaussian distribution $G_{\mathbf{Y}}$ with mean $\mu_{\mathbf{Y}} = (\mu, \mu)'$ and covariance structure

$$\Sigma(\rho) = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

For the forecasting model, we also consider a bivariate Gaussian distribution $F_{\mathbf{X}}$. In the subsequent simulation study, we consider the following differences between the distributions $G_{\mathbf{Y}}$ and $F_{\mathbf{X}}$:

- (*Error in mean*) The forecasting distribution $F_{\mathbf{X}}$ is given by a bivariate Gaussian distribution with correct covariance structure $\Sigma(\rho)$ but a misspecified mean $\mu_{\mathbf{X}} = (\hat{\mu}, \hat{\mu})'$.
  This means for every replication the forecaster iusses a forecast describing the multivariate distribution of a random variable X such that

  $$X \sim \mathcal{N}_2(\mu_{\mathbf{X}}, \Sigma(\rho)).$$

  The resulting difference between $G_{\mathbf{Y}}$ and $F_{\mathbf{X}}$ is pictured in Figure 7.1a.

- (*Error in variance*) Here only the variance is misspecified, i.e. for every replication the forecaster iusses a forecast describing the multivariate distribution of the random variable $\mathbf{X}$ such that

  $$\mathbf{X} \sim \mathcal{N}_2(\mu_{\mathbf{Y}}, \hat{\Sigma}(\rho)),$$

  where

  $$\hat{\Sigma}(\rho) = \hat{\sigma}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

  For the resulting difference between the distributions $G_{\mathbf{Y}}$ and $F_{\mathbf{X}}$ see Figure 7.1b.

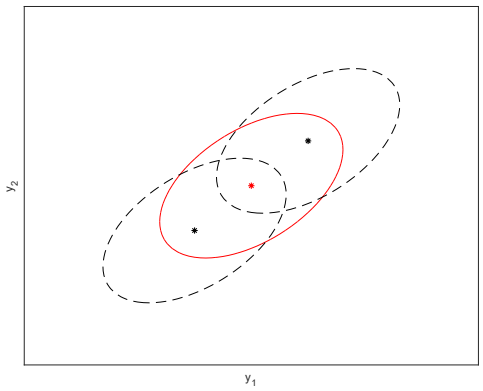
- (*Error in correlation*) In this case the forecaster makes an error when specifying the dependency structure, while describing the mean $\mu_{\mathbf{Y}}$ and the variance σ^2 of the process correctly. This means for every experiment the forecaster iusses a forecast describing the multivariate distribution of a random variable $\mathbf{X}$ such that

  $$\mathbf{X} \sim \mathcal{N}_2(\mu_{\mathbf{Y}}, \Sigma(\hat{\rho})),$$

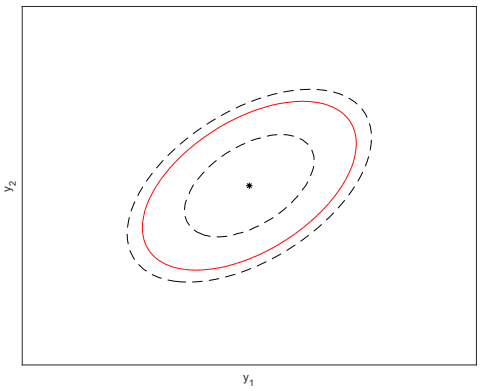
  where

  $$\Sigma(\hat{\rho}) = \sigma^2 \begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}.$$

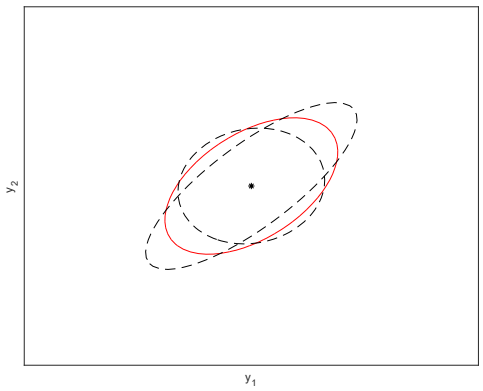
  The resulting difference between the distributions is pictured in Figure 7.1c.



(a) errors in mean



(b) errors in variance



(c) errors in correlation

Figure 7.1.: Illustration of different misspecifications in the process generating distribution. Each distribution is represented by a single isoline of its corresponding density. The volume over the area bounded by the ellipse equals α. The solid red lines represent the density generating the real process, the dotted black lines the misspecified densities. In (a) errors in mean are shown. The misspecified densities are shifted copies of the real density in direction of the major axis. In (b) errors in variance are shown. This results in inflation ($\hat{\sigma}^2 > \sigma^2$) respectively deflation ($\hat{\sigma}^2 < \sigma^2$) of the ellipse. In (c) errors in correlation are shown. This results in stretching the ellipse in direction of its major axis ($\hat{\rho} > \rho$) or in direction of its minor axis ($\hat{\rho} < \rho$).

To evaluate the forecasting distributions we utilize the following multivariate scoring rules:

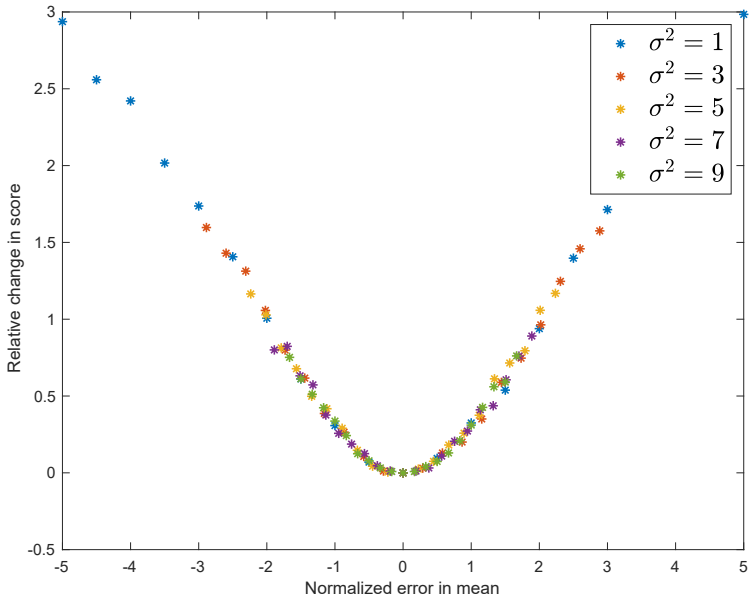
- Energy score $\mathrm{ES}_{0.5}$, estimated by $\widehat{\mathrm{ES}}_{0.5}^{band}$,
- Energy score ES_{1}, estimated by $\widehat{\mathrm{ES}}_{1}^{band}$,

- Energy score $\mathrm{ES}_{1.5}$, estimated by $\widehat{\mathrm{ES}}_{1.5}^{band}$,
- Variogram score $\mathrm{VS}_{W,1}$, estimated by $\widehat{\mathrm{VS}}_{W,1}$, with $W = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$,
- Dawid-Sebastiani score DSS, estimated by $\widehat{\mathrm{DSS}}$.

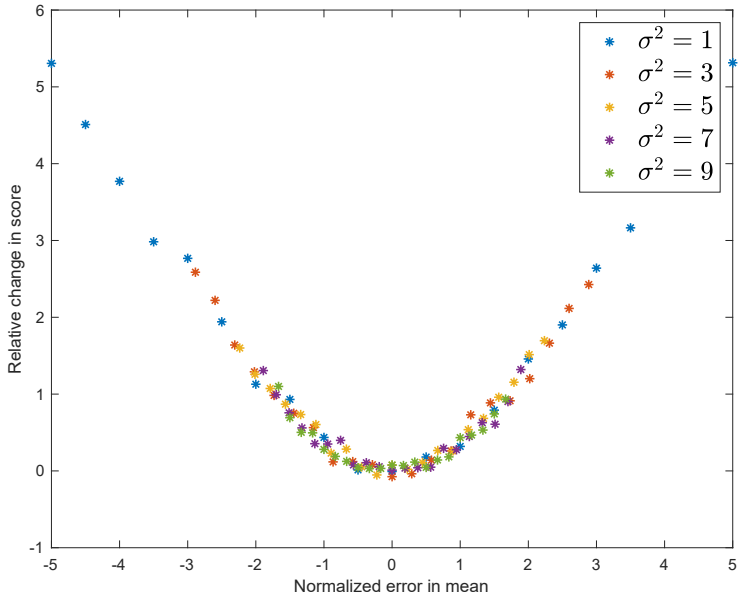
As already stated above, in this simulation study we compute the relative change in score and the DM-test statistic both with respect to the true model.

7.3.1. Errors in mean

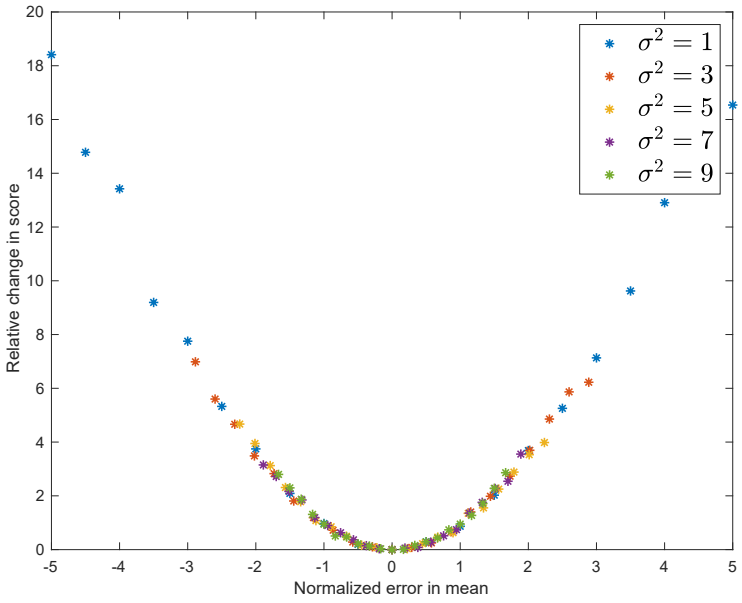
First, we consider errors in mean: Let the real distribution Y be given by a bivariate Gaussian process with mean $\mu = 5$, correlation $\rho = 0.5$, and variance σ^2, where $\sigma^2 \in \{1, 3, 5, 7, 9\}$. In order to characterize the sensitivity to the process variance the following assessment is performed for different values of σ^2.
Note that any other values of μ and ρ would yield qualitatively similar results.
As we aim to determine the sensitivity to a misspecified mean we assume the correlation and the variance are reported correctly.
So let $\hat{\mu}$ be the mean parameter of the predictive distribution. We choose $\hat{\mu}$ out of the grid $\{0, 0.5, 1, 1.5, \ldots, 9.5, 10\}$. The relative change in score value ΔSc is evaluated as a function of the normalized error in mean $(\mu - \hat{\mu})/\sigma$. First of all, the energy score and the Dawid-Sebastiani score are able to identify the correctly specified forecasting distribution.
Furthermore, it can be noted that in particular for the energy score the relative prediction error seems independently of the process variance.
Since the energy score is based on an Euclidean distance the relative change in energy score solely depends on the magnitude of the normalized error in mean and not on the direction of the shift of the distribution along the translation axis, see Figure 7.1a. This means the discrimination ability is symmetric with respect to errors in mean.
Obvioulsy, the scale of the relative change in score value changes due to a different coefficient β, see Figure 7.2a, Figure 7.2b and Figure 7.2c.



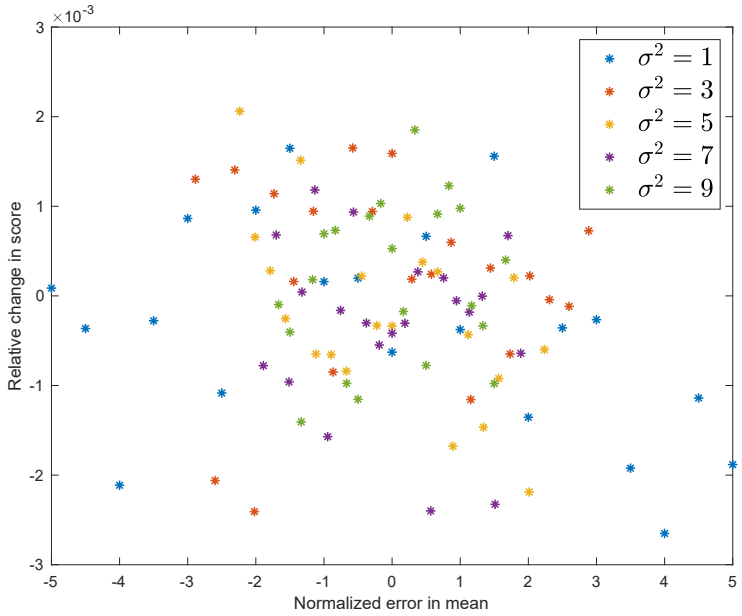
(a) $\Delta\text{ES}_{0.5}$



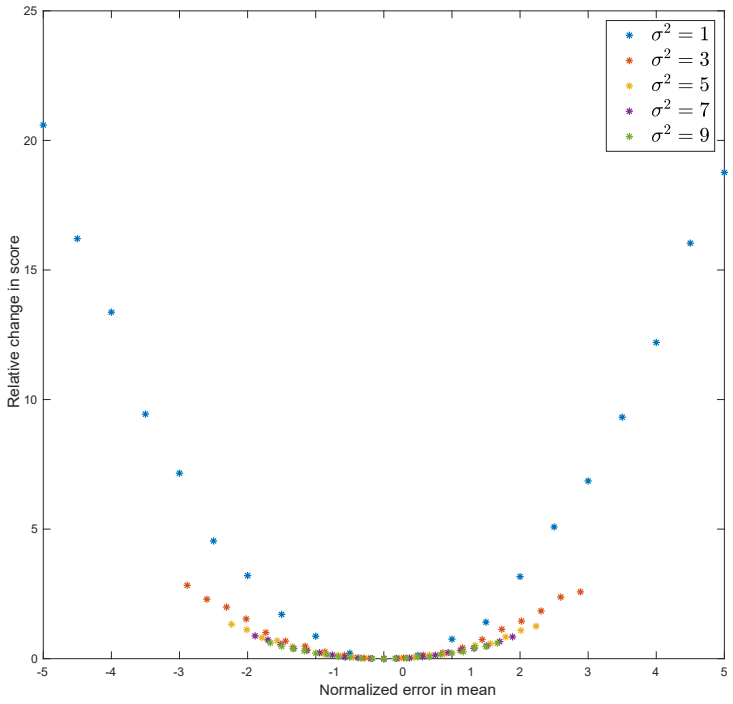
(b) ΔES_1



(c) $\Delta\mathrm{ES}_{1.5}$



(d) $\Delta\mathrm{VS}_{W,1}$



(e) Δ DSS

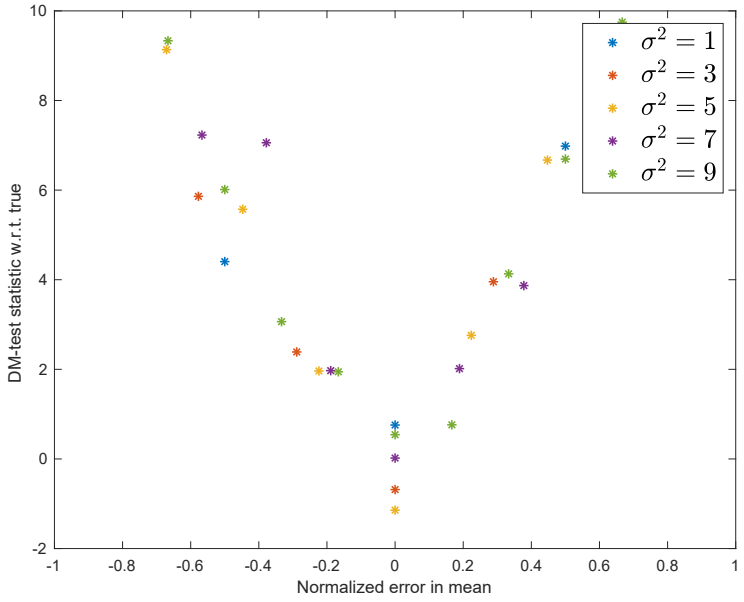
Figure 7.2.: Discrimination ability of the energy score (a), (b), and (c), the variogram score (d), and the Dawid-Sebastiani score (e) assessed with ΔSc, in terms of their sensitivity to prediction errors in mean for bivariate Gaussian predictive densities.

The variogram score isn't able to discriminate between the different distributions at all, see Figure 7.2d, as the bias in mean is the same in every component, and, therefore, cancels out. This follows directly from the definition of the variogram score.
Similar to the energy score, the Dawid-Sebastiani score is symmetric with respect to errors in mean. This follows directly from the Definition (3.26) of the Dawid-Sebastiani score.

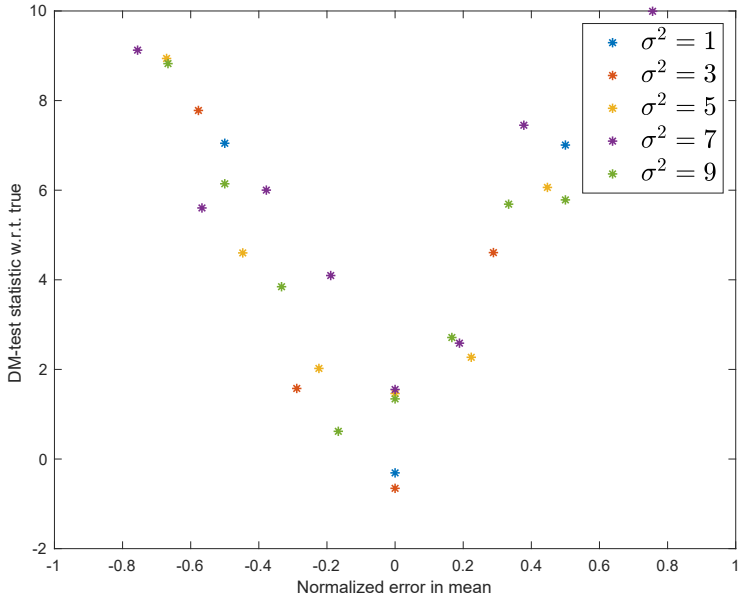
To statistically evaluate these results we apply the Diebold-Mariano test and compute the test statistic values corresponding to each forecasting experiment. Remember that the null hypothesis H_0 of equal predictive accuracy of two competing distributions is rejected at significance level α if the absolute value of the test statistic is greater than the $(1 - \alpha/2)$ - quantile of the standard normal distribution $z_{1-\alpha/2}$. A common choice is $\alpha = 5\,\%$ which corresponds to $z_{1-\alpha/2} = 1.9600$.

First, we note that all scoring rules except the variogram score are capable of detecting the calibrated forecast, see Figure 7.3. The Dawid-Sebastiani score and the

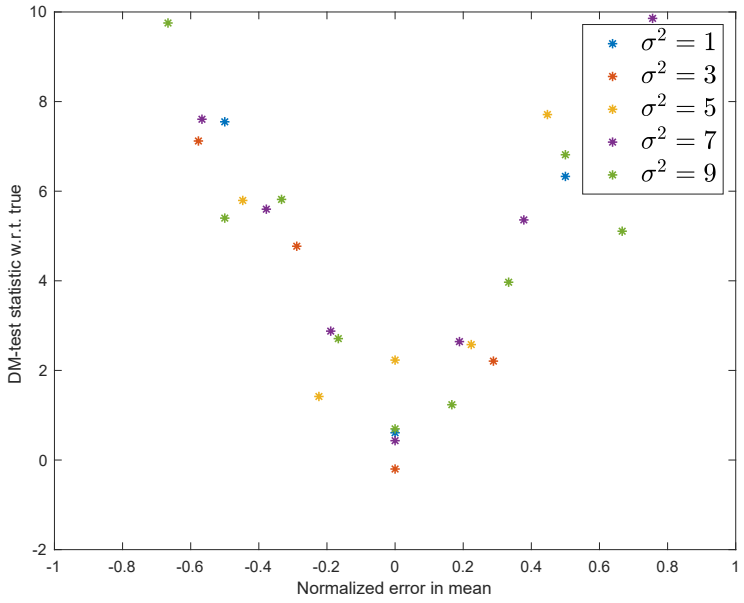
energy score yield large values for the test statistics and are comparable sensitive with respect to erros in mean. The energy score falsely accepts the null hypothesis of equal predictive performance solely for a normalized error of 0.1667 and the Dawid-Sebastiani score only for -0.1667.



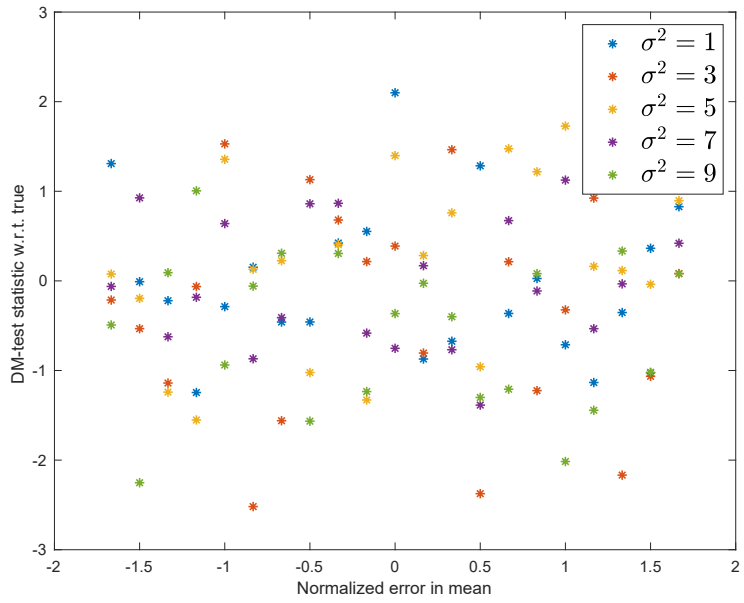
(a) $\mathrm{ES}_{0.5}$



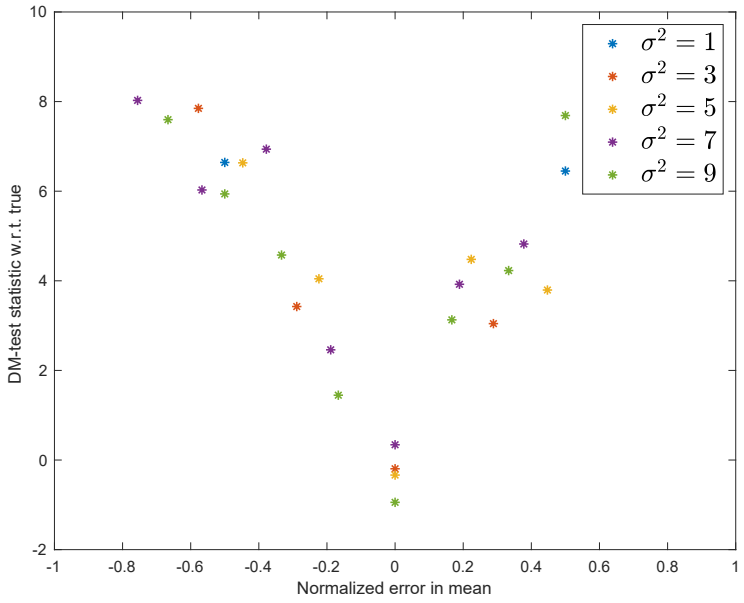
(b) ES_1



(c) $\mathrm{ES}_{1.5}$



(d) $\mathrm{VS}_{W,1}$



(e) DSS

Figure 7.3.: Discrimination ability of the energy score (a), (b), and (c), the variogram score (d), and the Dawid-Sebastiani score (e) assessed with the Diebold-Mariano test with respect to the true underlying distribution, in terms of their sensitivity to prediction errors in mean for bivariate Gaussian predictive densities. The statistics are capped at 10 to improve interpretability.

The energy score with greater parameter β falsely accepts the null hypothesis for a broader range of normalized errors in mean. This means the discrimination ability of the energy score decreases with an increasing parameter β.
As considering the relative change in score value indicates, nearly the DM-test statistic values of all forecasts evaluated with the variogram score fall in the range from -1.96 to 1.96, i.e. the variogram score is not able to discriminate between the different distributions.

Remark 7.3.1. As previously noted, it can be concluded from the simulation studies that the discrimination ability of the energy score with respect to errors in mean is independent of the process variance σ^2.
To be more precise, the relative change in score and also the values of the DM-test statistic seem to depend solely on the relative error in mean $(\mu - \hat{\mu})/\sigma$.
This is proven as followes:
Remember that the forecasting distribution is given by $\mathbf{X} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma(\rho))$, where $\mu_{\mathbf{X}} = (\hat{\mu}, \hat{\mu})'$ and

$$\Sigma(\rho) = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The true underlying distribution is given by $\mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{Y}}, \Sigma(\rho))$, where $\mu_{\mathbf{Y}} = (\mu, \mu)'$.

We can rewrite $\mathbf{X} = \mu_{\mathbf{X}} + \sigma \cdot \mathbf{U}$ and $\mathbf{Y} = \mu_{\mathbf{Y}} + \sigma \cdot \mathbf{U}$ respectively, where

$$\mathbf{U} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

The expected relative change in score for the energy score is given by

$$\text{RelChange}(F, G) := \frac{\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta}{\frac{1}{2}\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta}.$$

We can write these term as

$$\begin{aligned}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta &= \mathbb{E}\|\mu_{\mathbf{X}} + \sigma \cdot \mathbf{U} - \mu_{\mathbf{X}} + \sigma \cdot \mathbf{V}\|^\beta \\ &= \sigma^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta &= \mathbb{E}\|\mu_{\mathbf{Y}} + \sigma \cdot \mathbf{U} - \mu_{\mathbf{Y}} + \sigma \cdot \mathbf{V}\|^\beta \\ &= \sigma^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta &= \mathbb{E}\|\mu_{\mathbf{X}} + \sigma \cdot \mathbf{U} - \mu_{\mathbf{Y}} + \sigma \cdot \mathbf{V}\|^\beta \\ &= \sigma^\beta \mathbb{E}\left\|\frac{1}{\sigma} \cdot (\mu_{\mathbf{X}} - \mu_{\mathbf{Y}}) + \mathbf{U} - \mathbf{V}\right\|^\beta,\end{aligned}$$

where $\mathbf{U}, \mathbf{V}$ are independent, and

$$\mathbf{U}, \mathbf{V} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

It follows that:

$$\text{RelChange}(F, G) := \frac{\mathbb{E}\|\frac{1}{\sigma} \cdot (\mu_{\mathbf{X}} - \mu_{\mathbf{Y}}) + \mathbf{U} - \mathbf{V}\|^\beta - \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta}{\frac{1}{2}\mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta}.$$

Hence, we can conclude that the relative change in score does not depend on the process variance, but on the normalized error in mean $(\mu - \hat{\mu})(\sigma)$.

This calculation can analogously be carried out for the Diebold-Mariano test. Remember that $\overline{d}$ denotes the sample mean of the loss differential.
We now consider the closed-form counterpart, i.e. the expected loss differential $d := \text{Sc}(F^{(1)}, G) - \text{Sc}(F^{(2)}, G)$. In the current setting this yields

$$\begin{aligned}d &= \text{ES}(F, G) - \text{ES}(G, G) \\ &= \mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta \\ &= \sigma^\beta \mathbb{E}\|\frac{1}{\sigma}(\mu_{\mathbf{X}} - \mu_{\mathbf{Y}}) + \mathbf{U} - \mathbf{V}\|^\beta - \sigma^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta,\end{aligned}$$

where $\mathbf{U}, \mathbf{V}$ are independent, and

$$\mathbf{U}, \mathbf{V} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

Furthermore, it holds true for the variance of the loss differential that

$$\begin{aligned}&\mathrm{Var}\left(\|\mathbf{X}-\mathbf{Y}\|^\beta - \frac{1}{2}\|\mathbf{X}-\tilde{\mathbf{X}}\|^\beta - \frac{1}{2}\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^\beta\right)\\ &= \mathrm{Var}\left(\|\mathbf{X}-\mathbf{Y}\|^\beta\right) - \frac{1}{4}\mathrm{Var}\left(\|\mathbf{X}-\tilde{\mathbf{X}}\|^\beta\right) - \mathrm{Var}\left(\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^\beta\right).\end{aligned}$$

We observe that the single terms can be rewritten as

$$\begin{aligned}\mathrm{Var}\left(\|\mathbf{X}-\mathbf{Y}\|^\beta\right) &= \mathrm{Var}\left(\|\mu_\mathbf{X} + \sigma\cdot\mathbf{U} - \mu_\mathbf{Y} + \sigma\cdot\mathbf{V}\|^\beta\right)\\ &= \sigma^{2\beta}\mathrm{Var}\left(\left\|\frac{1}{\sigma}(\mu_\mathbf{X}-\mu_\mathbf{Y}) + \mathbf{U} - \mathbf{V}\right\|^\beta\right),\end{aligned}$$

$$\mathrm{Var}\left(\|\mathbf{X}-\tilde{\mathbf{X}}\|^\beta\right) = \sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right),$$

and

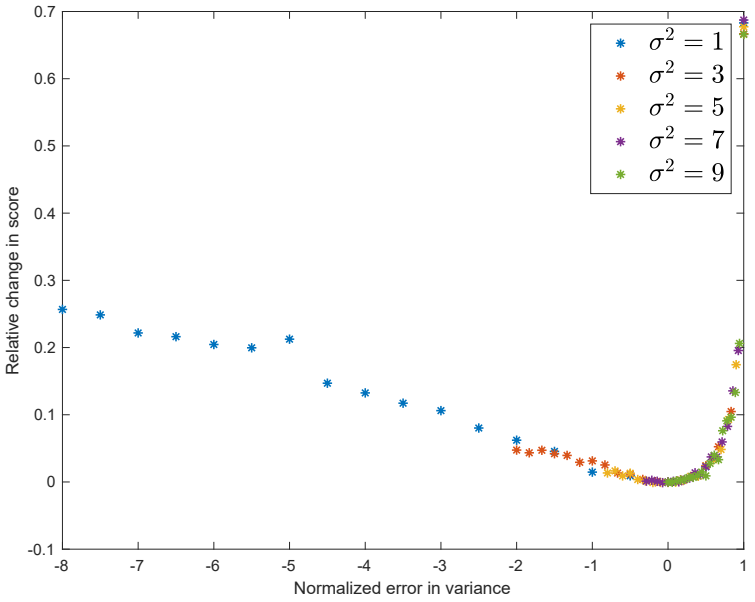
$$\mathrm{Var}\left(\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^\beta\right) = \sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right),$$

Therefore, the term σ^β cancels out, and, thus, the Diebold-Mariano test statistic depends on the normalized error in mean independent of the process variance.

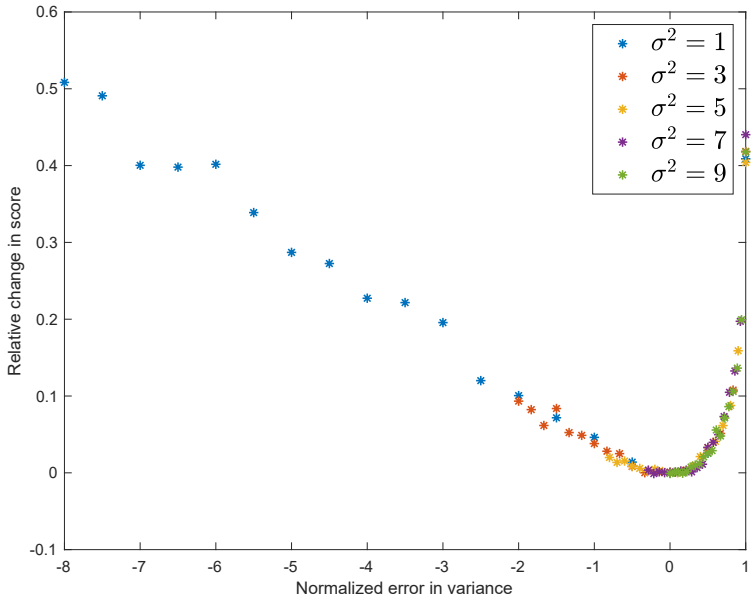
7.3.2. Errors in variance

The setup for errors in variance is similar: We keep the mean and correlation parameters fixed as $\mu_\mathbf{X} = \mu_\mathbf{Y} = (0,0)'$ and $\rho = 0.5$. In this example, we evaluate ΔSc as a function of the relative prediction error in variance which is defined as $(\sigma^2 - \hat{\sigma}^2)/\sigma^2$, where $\hat{\sigma}^2$ is the predictive variance.
We choose $\hat{\sigma}^2$ out of the grid $\{0, 0.5, 1, \ldots, 9, 5, 10\}$. As above, the assessment is performed for a set of σ^2 with $\sigma^2 \in \{1, 3, 5, 9\}$ to characterize the sensitivity to the process variance.
First, it can be stated that all three scoring rules are able to identify the calibrated forecast. All curves are zero or close to zero if the normalized error in variance is zero. Further, one can observe that the relative change in score value does not depend on the process variance.
In contrast to errors in mean, the discrimination ability of all three scoring rules is not symmetric. The relative change in score value ΔSc increases much steeper for too sharp densities than for densities that are too wide.

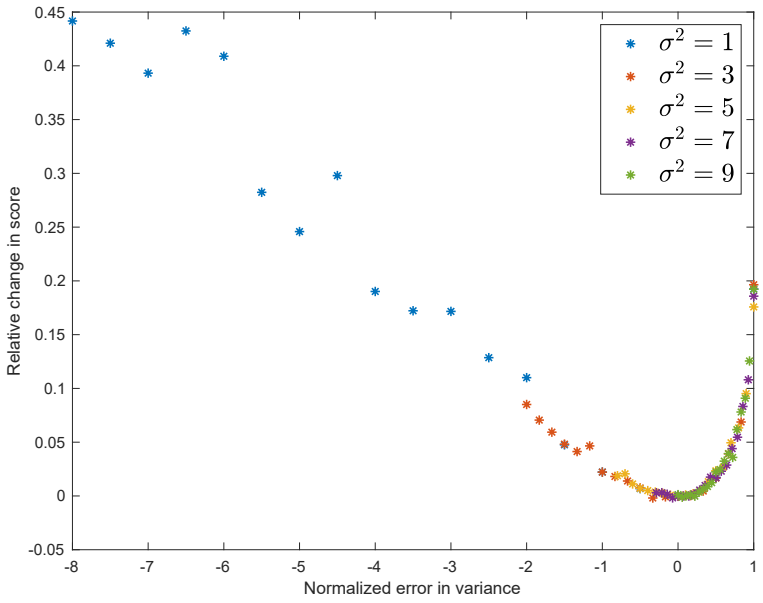
One can confirm the statement of Pinson and Tastu (2013) that the scales of the relative change in energy score in Figure 7.2a and Figure 7.4a, Figure 7.2b and Figure 7.4b such as Figure 7.2c and Figure 7.4c differ by a factor of $20 - 30$, and the relative changes in Dawid-Sebastiani score in Figure 7.2e and Figure 7.4e differ by a factor of 10.However, note that in this simulation study we compute the relative change in score value as a function of the normalized error in variance in contrast to the previous study. Therefore, this statement of Pinson and Tastu (2013) is initially not so meaningful.



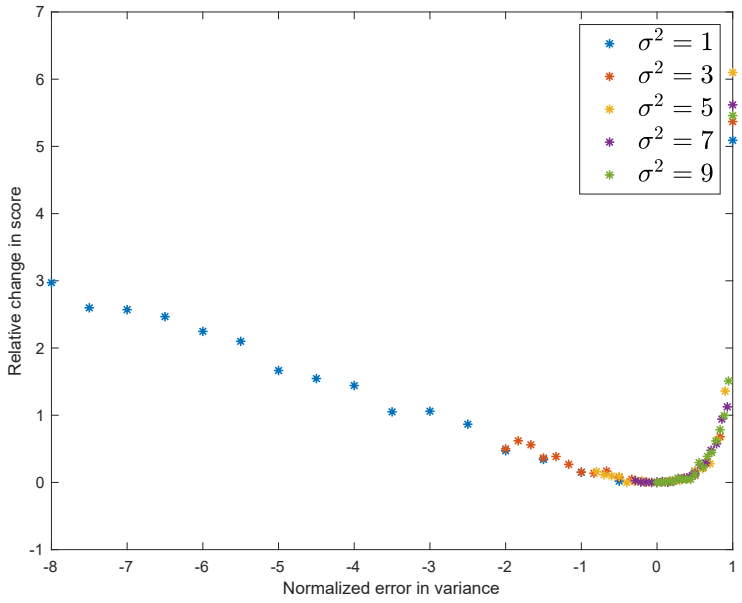
(a) $\Delta \mathrm{ES}_{0.5}$



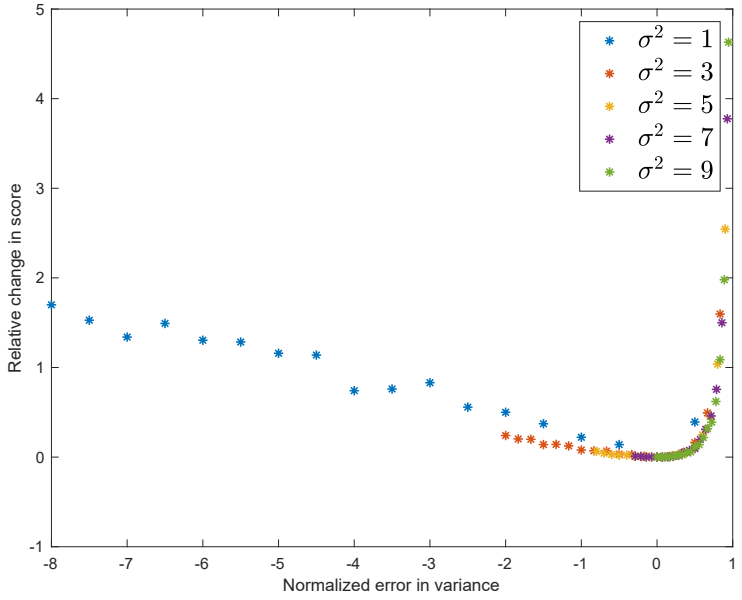
(b) ΔES_1



(c) $\Delta\text{ES}_{1.5}$



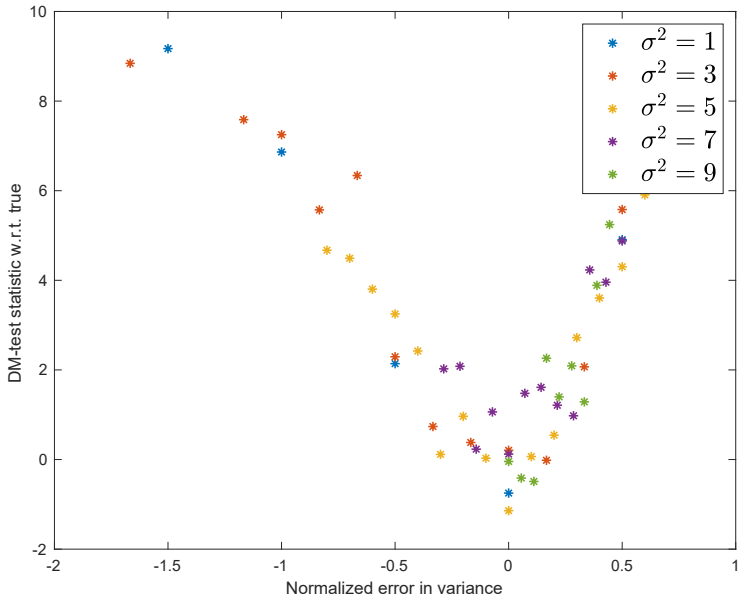
(d) ΔVS



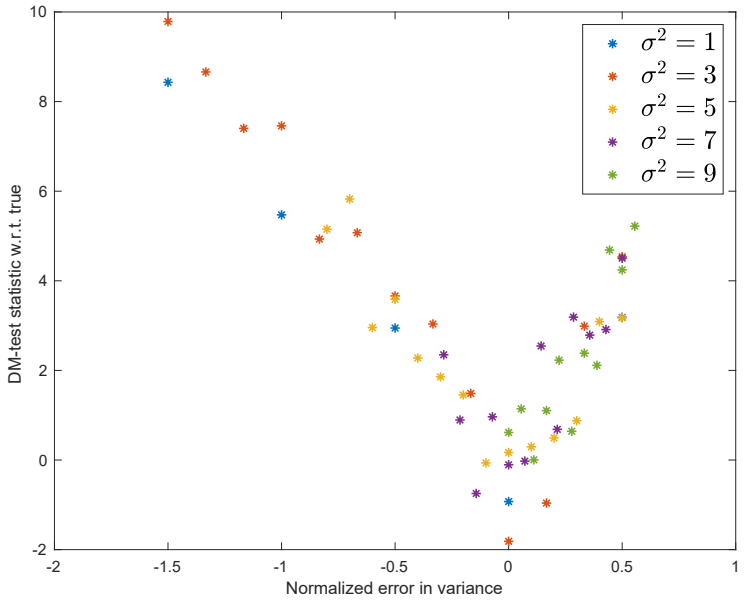
(e) ΔDSS

Figure 7.4.: Discrimination ability of the energy score (a), (b), and (c), the variogram score (d), and the Dawid-Sebastiani score (e) assessed with ΔSc, in terms of their sensitivity to prediction errors in variance for bivariate Gaussian predictive densities.

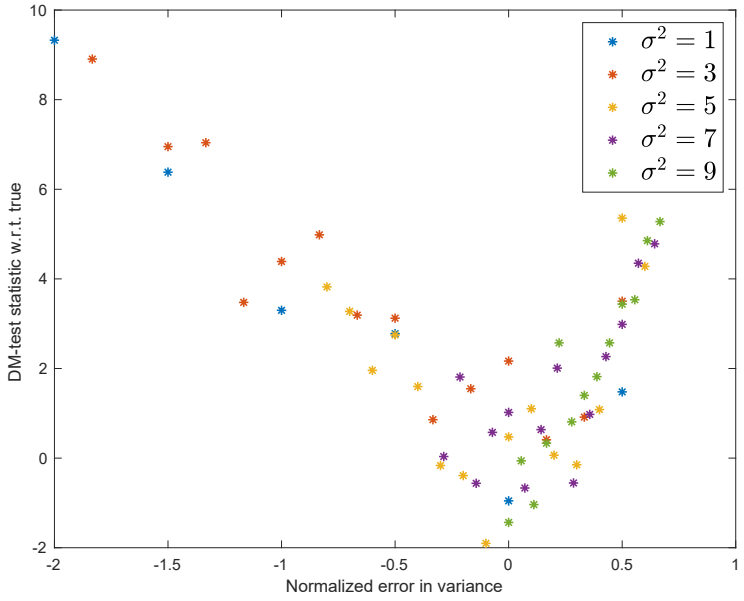
Afterwards, we apply the Diebold-Mariano test. As above, the statistics in Figure 7.5 are capped at 10. In this setting, the Dawid-Sebastiani score seems to have the best discrimination abilities. For the calibrated forecast all values of the DM-test statistic lie in the interval from -1.96 to 1.96, i.e. the null hypothesis is not falsely rejected at significance level $\alpha = 5\,\%$. This also applies to the energy score with coefficients $\beta = 0.5$ and $\beta = 1$ and the variogram score. However, it can be noticed that in general the values of the DM-test statistic are greatest in case of the Dawid-Sebastiani score followed by the energy score with coefficient $\beta = 0.5$. The values of the DM-test statistic for the energy score with $\beta = 1$ and the variogram score are smaller. Both seem to perform quite comparable in this simulation study. Lastly, we have the smallest values for the enery score with $\beta = 1.5$. In other words, the interval of normalized errors in mean for which the null hypothesis can (partially falsely) not be rejected is smallest for the Dawid-Sebastiani score followed by the energy score with parameter $\beta = 0.5$. So again, the Dawid-Sebastiani score has the best discrimination ability among the considered scoring rules.



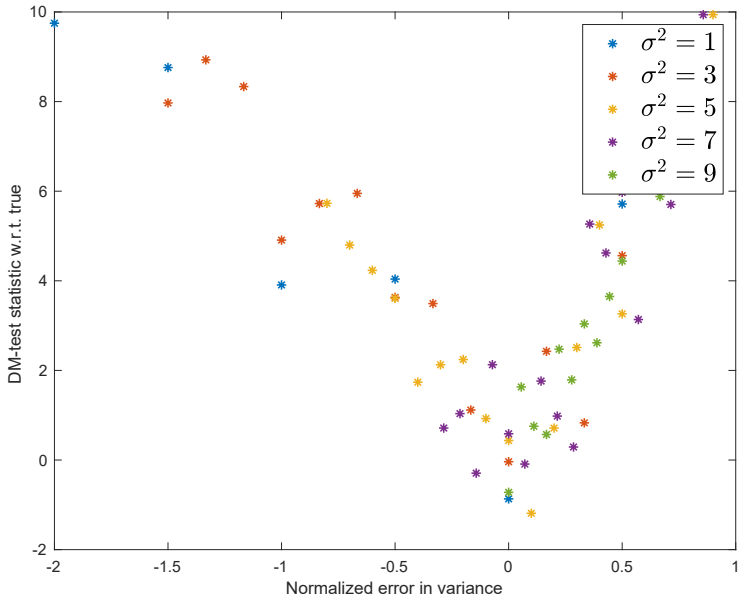
(a) $\mathrm{ES}_{0.5}$



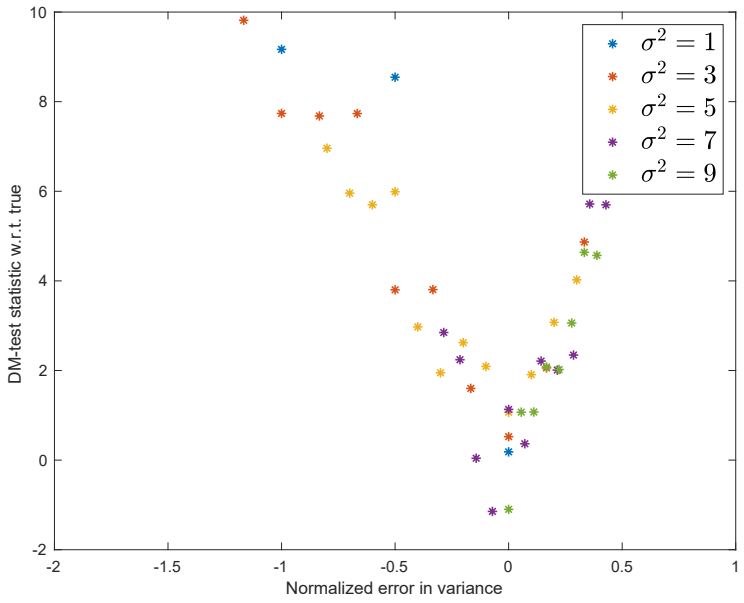
(b) ES_1



(c) $\mathrm{ES}_{1.5}$



(d) VS



(e) DSS

Figure 7.5.: Discrimination ability of the energy score (a), (b), and (c), the variogram score (d), and the Dawid-Sebastiani score (e) assessed with the Diebold-Mariano test, in terms of their sensitivity to prediction errors in variance for bivariate Gaussian predictive densities.

Remark 7.3.2. Similarly to the previous subsection, we can show that the relative change in score and the values of the Diebold-Mariano test statistic as a function of the normalized error in variance are independent of the process variance σ.
So let the forecasting distribution be given by $\mathbf{X} \sim \mathcal{N}\left(\mu_{\mathbf{Y}}, \hat{\Sigma}(\rho)\right)$ and the true underlying distribution is $\mathbf{Y} \sim (\mu_{\mathbf{Y}}, \Sigma(\rho))$, where

$$\Sigma(\rho) = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{, and} \quad \hat{\Sigma}(\rho) = \hat{\sigma}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Again, we can rewrite $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{X} = \hat{\sigma} \cdot \mathbf{U}$ and $\mathbf{Y} = \sigma \cdot \mathbf{U}$ respectively, where

$$\mathbf{U} \sim \mathcal{N}\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

First, note that the normalized error in variance can be rewritten as

$$\frac{\sigma^2 - \hat{\sigma}^2}{\sigma^2} = 1 - \frac{\hat{\sigma}^2}{\sigma^2}.$$

Hence, the relative change in score and the DM-test statistic values depend on the ratio of $\hat{\sigma}$ to σ. It holds for the expected relative change in score that

$$\text{RelChange}(F, G) \quad = \quad \frac{\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta}{\frac{1}{2}\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta},$$

where

$$\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta = \hat{\sigma}^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta,$$

$$\mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta = \sigma^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta,$$

and

$$\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta = \sigma^\beta \mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U} - \mathbf{V}\|^\beta,$$

where $\mathbf{U}, \mathbf{V}$ are independent, and

$$\mathbf{U}, \mathbf{V} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Therefore,

$$\begin{aligned} \text{RelChange}(F, G) \quad &= \quad = \frac{\sigma^\beta \mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U} - \mathbf{V}\|^\beta - \frac{1}{2}\hat{\sigma}^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta - \frac{1}{2}\sigma^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta}{\frac{1}{2}\sigma^\beta \mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta} \\ &= \quad \frac{2\mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U} - \mathbf{V}\|^\beta}{\mathbb{E}\|\mathbf{U} - \mathbf{V}\|^\beta} - \frac{\hat{\sigma}^\beta}{\sigma^\beta} - 1. \end{aligned}$$

So, the expected relative change in score depends on the the ratio $\hat{\sigma}/\sigma$. Next, we consider the Diebold-Mariano test and calculated the expected loss differential $d = \mathrm{Sc}(F^{(1)}, G) - \mathrm{Sc}(F^{(2)}, G)$. It holds that

$$\begin{aligned} d &= \mathbb{E}\|\mathbf{X}-\mathbf{Y}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{X}-\tilde{\mathbf{X}}\| - \frac{1}{2}\mathbb{E}\|\mathbf{Y}-\tilde{\mathbf{Y}}\| \\ &= \sigma^\beta \mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^\beta - \frac{1}{2}\hat{\sigma}^\beta\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta - \frac{1}{2}\sigma^\beta\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta \\ &= \sigma^\beta \mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta\left(\hat{\sigma}^\beta+\sigma^\beta\right), \end{aligned}$$

where $\mathbf{U}, \mathbf{V}$ are independent. Moreover, for the variance of the loss differential it holds that

$$\begin{aligned} &\mathrm{Var}\left(\|\mathbf{X}-\mathbf{Y}\|^\beta - \frac{1}{2}\|\mathbf{X}-\tilde{\mathbf{X}}\|^\beta - \frac{1}{2}\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^\beta\right) \\ &= \sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\frac{\hat{\sigma}}{\sigma}\mathbf{V}\|^\beta\right) + \frac{1}{4}\hat{\sigma}^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right) + \frac{1}{4}\sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right). \end{aligned}$$

Let us now consider

$$\begin{aligned} &\frac{d^2}{\mathrm{Var}\left(\|\mathbf{X}-\mathbf{Y}\|^\beta - \frac{1}{2}\|\mathbf{X}-\tilde{\mathbf{X}}\|^\beta - \frac{1}{2}\|\mathbf{Y}-\tilde{\mathbf{Y}}\|^\beta\right)} \\ &= \frac{\left(\sigma^\beta\mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{U}-\mathrm{V}\|^\beta\left(\hat{\sigma}^\beta+\sigma^\beta\right)\right)^2}{\sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\frac{\hat{\sigma}}{\sigma}\mathbf{V}\|^\beta\right) + \frac{1}{4}\hat{\sigma}^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right) + \frac{1}{4}\sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right)} \\ &= \Big(\sigma^{2\beta}\mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^{2\beta} - (\sigma^{2\beta}+\hat{\sigma}\sigma)\mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^\beta\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta \\ &\quad + \frac{1}{4}\left(\hat{\sigma}^{2\beta}+2\sigma^\beta\hat{\sigma}^\beta+\sigma^{2\beta}\right)\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta\Big)\cdot\Big(\sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\frac{\hat{\sigma}}{\sigma}\mathbf{V}\|^\beta\right) \\ &\quad + \frac{1}{4}\hat{\sigma}^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right) + \frac{1}{4}\sigma^{2\beta}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right)\Big)^{-1} \\ &= \Bigg[\mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^{2\beta} - \left(1+\frac{\hat{\sigma}}{\sigma}\right)\mathbb{E}\|\frac{\hat{\sigma}}{\sigma}\mathbf{U}-\mathbf{V}\|^\beta\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta \\ &\quad + \frac{1}{4}\left(\frac{\hat{\sigma}^{2\beta}}{\sigma^{2\beta}}+2\frac{\hat{\sigma}^\beta}{\sigma}+1\right)\mathbb{E}\|\mathbf{U}-\mathbf{V}\|^\beta\Bigg]\cdot\Big(\mathrm{Var}\left(\|\mathbf{U}-\frac{\hat{\sigma}}{\sigma}\mathbf{V}\|^\beta\right) \\ &\quad + \frac{1}{4}\frac{\hat{\sigma}^{2\beta}}{\sigma^{2\beta}}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right) + \frac{1}{4}\mathrm{Var}\left(\|\mathbf{U}-\mathbf{V}\|^\beta\right)\Big)^{-1}. \end{aligned}$$

Therefore, the value of the DM-test statistic depends not on the process variance σ, but on the ratio $\hat{\sigma}/\sigma$ of the forecasted variance and the process variance.

7.3.3. Errors in correlation

Lastly, we consider the discrimination ability of the scoring rules in terms of errors in correlation: Here we keep the mean and variance parameters fixed as $\mu_{\mathbf{Y}} = (0,0)$

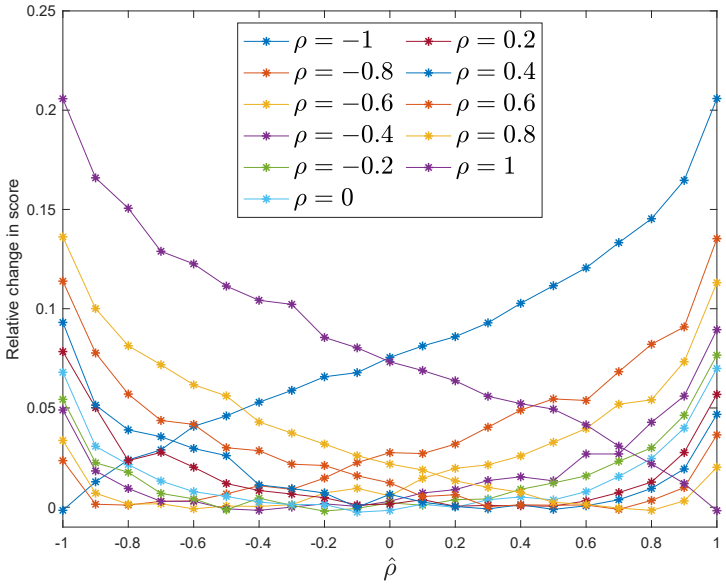
and $\sigma^2 = 1$, and calculate the relative change in score value ΔSc as a function of the predicted correlation $\hat{\rho}$. The assessment is performed for different values of

$$\rho \in \{-1, -0.8, \ldots, 0.8, 1\}.$$

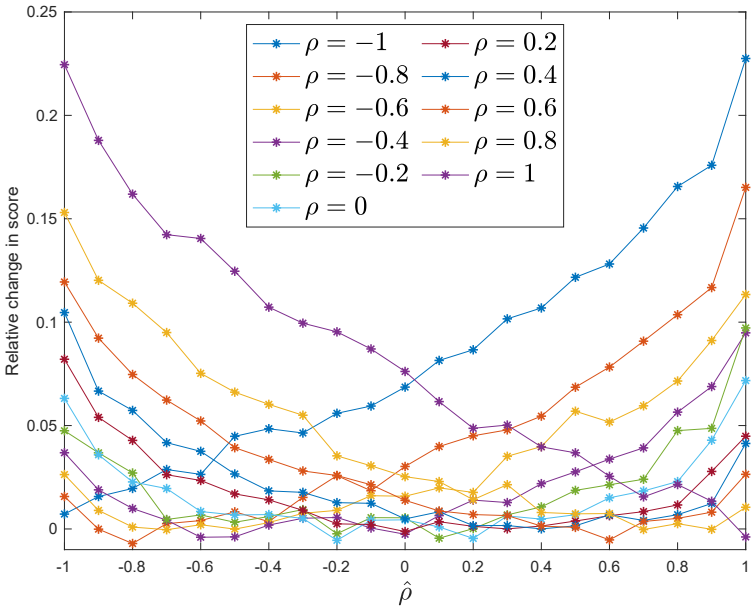
For the predicted correlation we choose the denser grid

$$\hat{\rho} \in \{-1, -0.9, -0.8, \ldots, 0.9, 1\}.$$

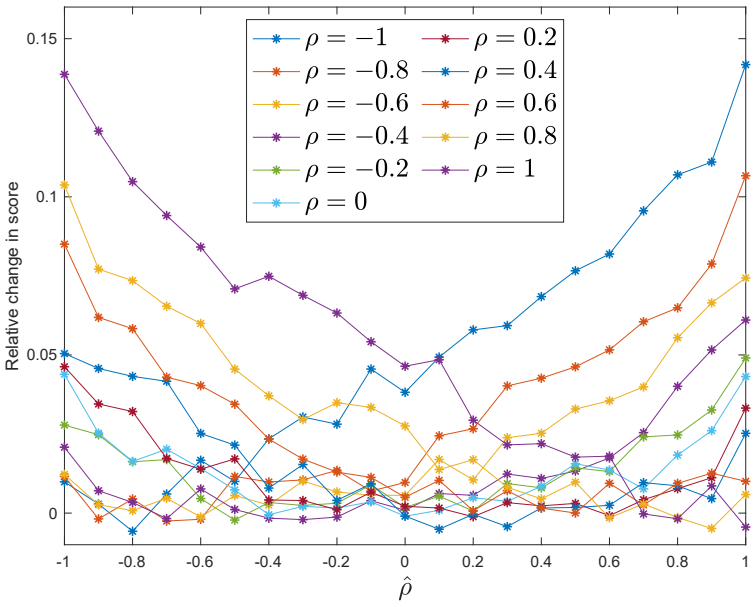
First of all, it can be noticed that the relative change in score value is quite small in comparison to the previous forecasting experiments. Further, it can be observed that the discrimination ability of the energy score, see Figure 7.6a, Figure 7.6b, and Figure 7.6c, and the Dawid-Sebastiani score, see Figure 7.6e, seems symmetric for positive and negative correlations, whereas the variogram score, see Figure 7.6d, is more sensitive for positive correlations. Regarding the energy score, one can note that the curve of the relative change in score value is smoother for a smaller coefficient β, see Figure 7.6a, Figure 7.6b, and Figure 7.6c.
The greatest value for the change in energy score is visually $\Delta\text{ES}_\beta \approx 0.20$ for $\beta = 0.5$ and $\beta = 1$. For $\beta = 1.5$ the relative change in score value is even smaller than 0.15.
These values are obtained in the extreme case, where the real process is perfectly correlated $\rho = 1$, and the forecast negatively correlated with $\hat{\rho} = -1$ and vice versa.
Obviously, these extreme cases are rather unusual in practice. So, for instance, let us consider the case where the true correlation is given by $\rho = 0.2$, and the forecasting correlation is $\hat{\rho} = 0.8$. For $\beta = 0.5$ and $\beta = 1$ the relative change in score is just approximately 2.5 %, for $\beta = 1.5$ even just approximately 1 %.
In general, it can be stated that the majority of values for the relative change in score value is within the range from 0 to 5 %.



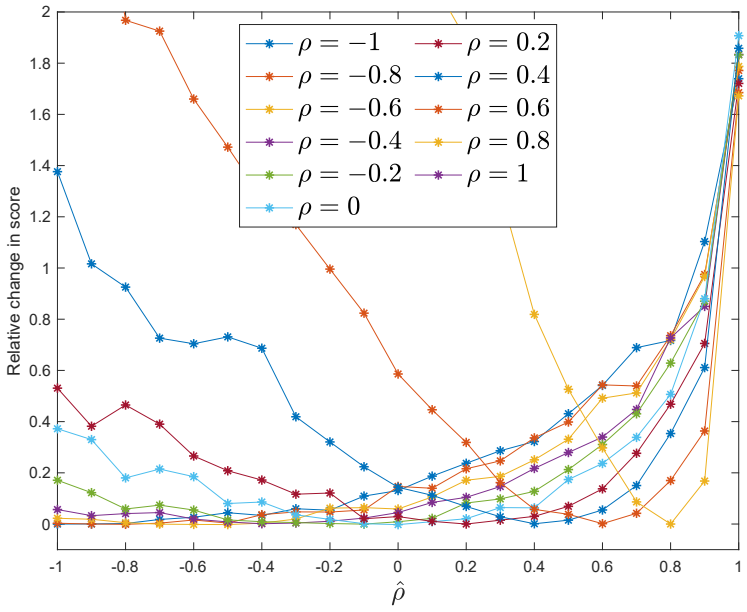
(a) $\Delta\text{ES}_{0.5}$



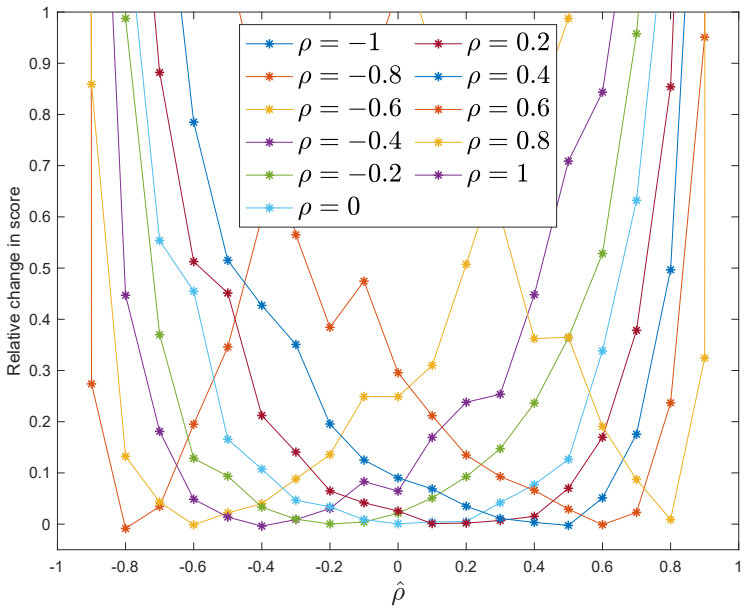
(b) ΔES_1



(c) $\Delta\text{ES}_{1.5}$



(d) ΔVS



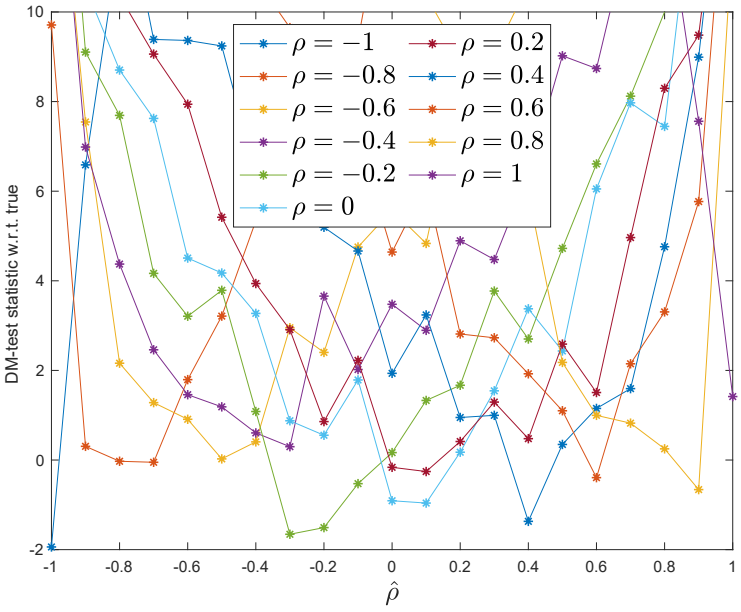
(e) ΔDSS

Figure 7.6.: Discrimination ability of the energy score (a), (b), and (c), the variogram score (d), and the Dawid-Sebastiani score (e) assessed with ΔSc, in terms of their sensitivity to prediction errors in correlation for bivariate Gaussian predictive densities. The process mean and variance parameters are kept fixed as $\mu = (0, 0)$ and $\sigma^2 = 1$.

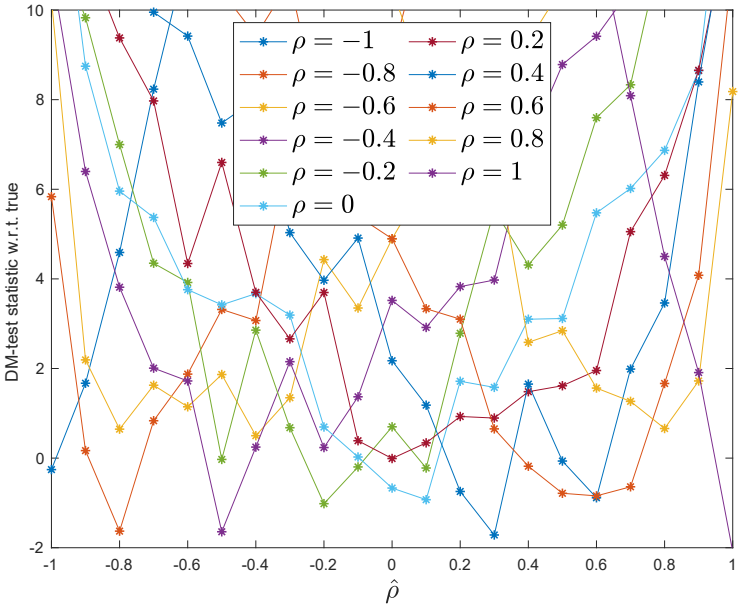
In the same situation as described above, where $\rho = 0.2$ and $\hat{\rho} = 0.8$, the relative change in variogram score is approximately 45 %. However, the situation of positive correlation of the underlying process and the forecast favors the variogram score. If the process correlation is given, for instance, by $\rho - 0.2$, and the forecasting correlation by $\hat{\rho} = -0.8$, the relative change in variogram score is approximately 1 %. Note that the relative change in score value for both negative process correlation and negative forecasting correlation is relatively low. Lastly, we consider the Dawid-Sebastiani score. Here the relative change in score value is approximately 80 % for $\rho = 0.2$ and $\hat{\rho} = 0.8$. Furthermore, the discrimination ability is symmetric for positive and negative values of ρ. The relative change in score value for $\rho = 0.2$ and $\hat{\rho} = 0.8$ is also approximately 80 % in contrast to the variogram score.

As in the preceding subsections, we apply the Diebold-Mariano test and calculate the values of the DM-test statistic as a function of the predicted correlation $\hat{\rho}$. The statistics are capped at 10 to improve interpretability.
We notice that the discrimination ability of the energy score is roughly symmetric for positive and negative correlations. Furthermore, one might infer that the discrimination ability of the energy score is better for a smaller coefficient β. This can be concluded by the fact that the interval of predicted correlations, in which the values of the DM-test statistic are smaller than 1.96, is shorter for a smaller coefficient β.
By computing the DM-test statistic, the asymmetry of the variogram score regarding positive and negative correlations of the underlying process becomes evident. For positive correlations the null-hypothesis of equal predictive accuracy is rejected with significance level 5 % for all forecasts except the calibrated one, i.e. the values of the DM-test statistic are greater than 1.96 for all wrongly predicted correlations.
However, this does not apply to negative correlations of the underlying process. As an example we consider $\rho = -0.8$. In this case, the DM-test statistic values are smaller than 1.96 for the forecasting distributions with $\hat{\rho} = -0.9, -0.8, -0.7$, and -0.6, i.e. the variogram score is not able to discriminate these distributions.

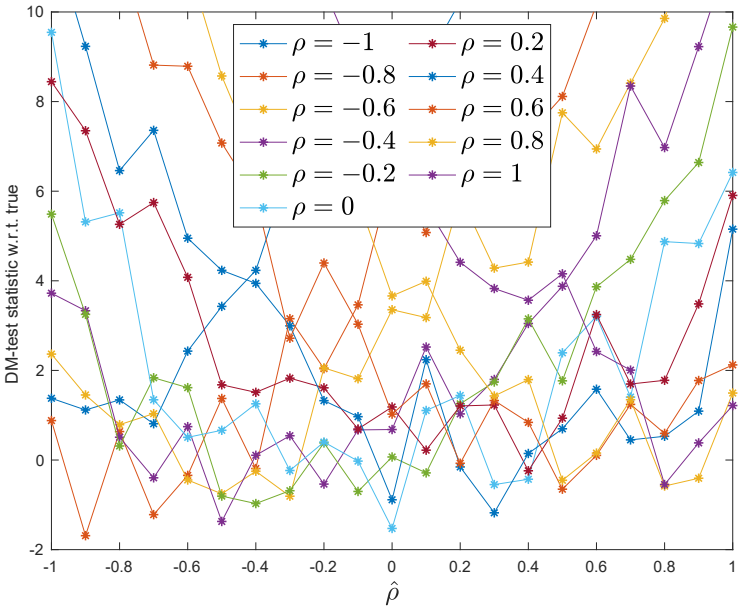
Lastly, we consider the Dawid-Sebastiani score. The discrimination ability of this scoring rule is symmetric for positive and negative process correlation. It appears that the Dawid-Sebastiani score is more sensitive for correlations with greater absolute values. For process correlation $\rho = -0.8$ and forecasted correlation $\hat{\rho} = -0.8$ the corresponding DM-test statistic is approximately 0, for $\hat{\rho} = -0.9$ the corresponding DM-test statistic is greater than 4 and for -0.7 even approximately 6. In contrast, the absolute values of the DM-test statistic corresponding to the forecasts with correlation $\hat{\rho} = 0.1, 0.2$ and 0.3 are smaller than 1.96 for a process with correlation $\rho = 0.2$, i.e. the Dawid-Sebastiani score is not able to discriminate between these distributions.



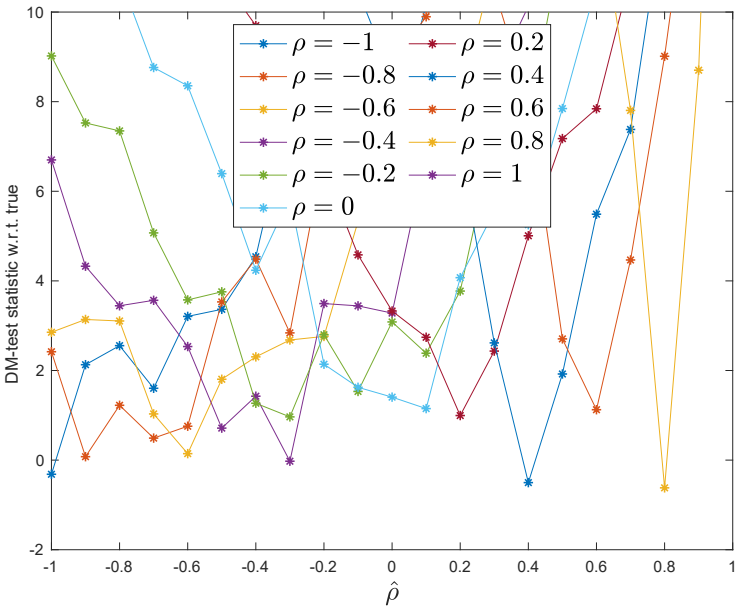
(a) $\mathrm{ES}_{0.5}$



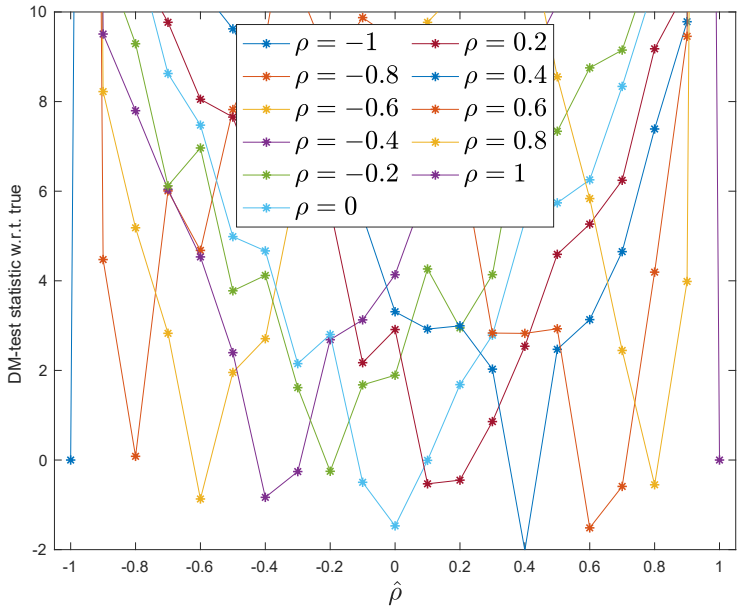
(b) ES_1



(c) $\mathrm{ES}_{1.5}$



(d) VS



(e) DSS

Figure 7.7.: Discrimination ability of the energy score (a), (b), and (c), the variogram score (d), and the Dawid-Sebastiani score (e) assessed with the Diebold-Mariano test, in terms of their sensitivity to prediction errors in correlation for bivariate Gaussian predictive densities. The process mean and variance parameters are kept fixed as $\mu_{\mathbf{Y}} = (0, 0)$ and $\sigma^2 = 1$.

7.3.4. Discussion of the study results

Considering the relative change in score our simulation study yields the same results as the study of Pinson and Tastu (2013). So, based on that the conclusion of the bad discrimination ability of the energy score with respect to errors in correlation can be verified.
The authors emphasize their conclusion of the bad discrimination ability by calculating an upper bound for the relative change in score for errors in correlation when the mean and variances of the process are predicted correctly. For that, it is assumed that the true distribution is given by the d-dimensional Gaussian distribution with zero mean and perfectly correlated components. That is $\mathbf{Y} \sim \mathcal{N}(0, \Sigma)$ with

$$\Sigma = \sigma^2 \mathbf{1}_{(d \times d)},$$

where $\mathbf{1}_{(d \times d)}$ is a $d \times d$-matrix of ones. The forecast is given by the naive forecast that totally neglects the interdependence structure between the single components, that is $\mathbf{X} \sim \mathcal{N}(0, \hat{\Sigma})$ with

$$\hat{\Sigma} = \sigma^2 \mathrm{diag}(\mathbf{1}_d),$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. In this thesis, we extend the calculation by the parameter β for the generalized energy score and are able to find a closed-form solution for the relative change in score for a general parameter β in this extreme case, see Appendix A.3.
It holds that

$$\begin{aligned}\Delta\mathrm{ES}_\beta(F) &= \frac{\mathrm{ES}_\beta(F,G) - \mathrm{ES}_\beta(G,G)}{\mathrm{ES}_\beta(G,G)} \\ &= 2^{1-\beta} n^{-\beta/2} \sqrt{\pi} \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{1+\beta}{2}\right)} \\ &\cdot \left(\frac{2^{\beta/2}}{\sqrt{d+1}}\, {}_2F_1\left(\frac{d+\beta}{2}, \frac{1}{2}, \frac{d}{2}; \frac{d}{d+1}\right) - 2^{\beta-1} \right) - 1,\end{aligned}$$

where ${}_2F_1$ denotes the Gauss hypergeometric function. As already stated in Pinson and Tastu (2013), this relative change in score is increasing in dimension size d, but reaches an asymptote of less than 15 % for higher dimensions if $\beta = 1$.
For a bivariate Gaussian distribution and $\beta = 1$ the upper bound is approximately 7.4 %. This upper bound is considerably less than for making errors in predicting the variance or mean parameter for the multivariate Gaussian process. Therefore, the authors conclude that these findings confirm the apparently poor discrimination ability of the energy score with respect to errors in correlation.
However, solely considering the relative change in score value is incomplete as the statistical significance of the results is not considered. Therefore, it is crucial to assess the score values with the Diebold-Mariano test.
The values of the corresponding DM-test statistic indicate that the energy score not only discriminates well among distributions with different mean and variance parameters, but also discriminates well between forecasts with different interdependence structure.
This clearly contradicts the conclusions of Pinson and Tastu (2013). Further, it can be noted that a smaller coefficient β leads to an improved discrimination ability of the energy score. The energy score with parameter $\beta = 0.5$ has an even better discrimination abilility than the variogram score which is specifically designed to detect errors in the interdependence structure.
Further, the energy score with parameter $\beta = 0.5$ and the Dawid-Sebastiani score perform quite similarly. However, note that the setting of this simulation study clearly favors the Dawid-Sebastiani score because we consider a Gaussian distribution.

Therefore, the conclusion of the poor discrimination ability of the energy score as stated e.g. by Pinson and Tastu (2013) has to be negated because the score values of miscalibrated forecasts differ statistically significant from the score values of the perfect forecast when the score values are assessed with the Diebold-Mariano test.
Lastly, we take a closer look at the influence of the parameter β on the discrimination ability of the energy score.

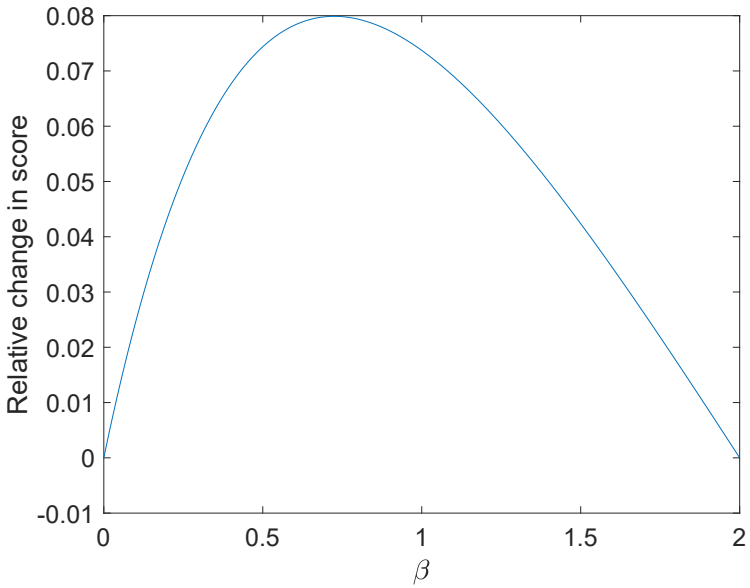


Figure 7.8.: Relative change in energy score as a function of β for a bivariate Gaussian process, when the forecast is given by the naive forecast and the true underlying distribution has perfectly correlated componentes. The mean and variance parameter are reported correctly.

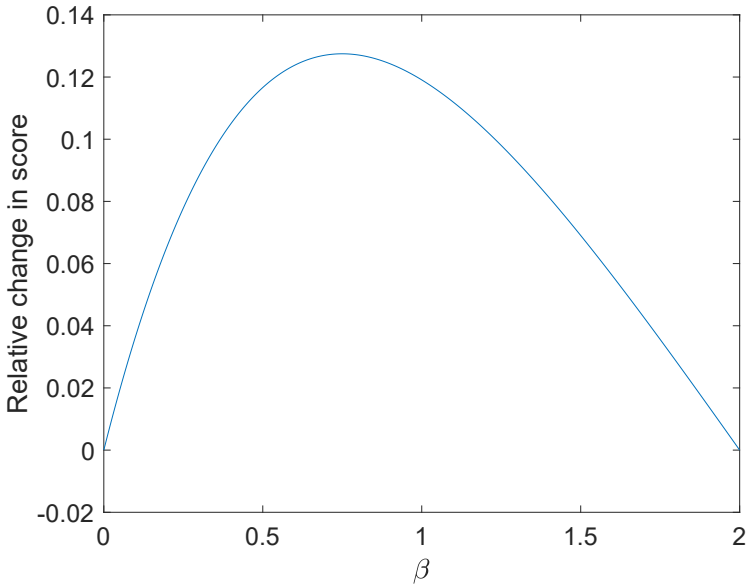


Figure 7.9.: Relative change in energy score as a function of β for a Gaussian process of dimension $d = 5$, when the forecast is given by the naive forecast and the true underlying distribution has perfectly correlated componentes. The mean and variance parameter are reported correctly.

For dimension $d = 2$, the relative change in score has the maximal value of approximately $7,99\,\%$ for $\beta \approx 0.7247$. For $\beta = 0$ and $\beta = 2$ the relative change in score is obviously zero, see Figure 7.3.4.
For dimension $d = 5$, the maximal relative change in score is approximately $12.75\,\%$ for $\beta \approx 0.7496$. So the parameter β, for which the relative change in score attains its maximum, depended on the dimension d of the process, but in both cases it is smaller than 1, see Figure 7.9.
However, considering the Diebold-Mariano test statistic values, see Figure 7.10, suggests that choosing β as small as possible is optimal as the DM-test statistic value is greater the smaller the parameter β is.
This contradicts the previous findings for which the relative change in score was considered.

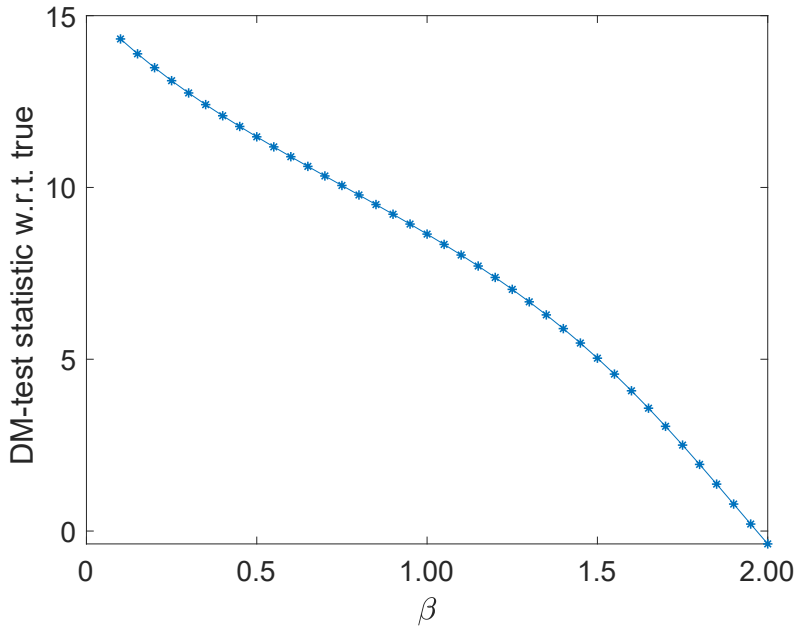


Figure 7.10.: DM-test statistic values as a function of β for a bivariate Gaussian process, when the forecast is given by the naive forecast and the true underlying distribution has perfectly correlated components. The mean and variance parameter are reported correctly.

7.4. Simulation study II

In the subsequent section we reproduce and extend the simulation study of Scheuerer and Hamill (2015). We compare the energy score with coefficients $\beta = 0.5, 1$, and 1.5, the Dawid-Sebastiani score and the variogram score of order $p = 0.1, 0.5, 1$, and 2. In contrast to the previous simulation study, we use inverse distance weights for the weight matrix of the variogram score.

In all experiments, we generate $N = 5000$ observations vectors of dimension d and a m-member ensemble of forecast vectors of the same dimension with both correct and misspecified means, variances or correlations.
In the previous simulation study a large ensemble ($m = 2^{14}$) was considered. In practice, this is rather unusual, particularly, in the field of weather forecasting. Therefore, we consider a small ($m = 20$) and medium-sized ($m = 100$) ensemble in this study. Furthermore, we aim to unterstand the impact of these different ensemble sizes on the different scores.
Each experiment is repeated 10 times und the respective outcomes are visualized by boxplots. This means we compute 10 score values for each probabilistic forecast.
Here we utilize another measurement for the discrimination ability of the different scoring rules adapted from Scheuerer and Hamill (2015). The authors utilize the overlapping of the boxes corresponding to different forecasting distributions as a measurement for the discrimination ability of a given scoring rule. If the boxes corresponding to different forecasting distributions do not overlap, this is interpreted as a good discrimination ability of the scoring rule considered. Contrary, if the boxes corresponding to different forecasts overlap, this is interpreted as a poor discrimination ability of the scoring rule.
However, we do not solely consider the score values of the forecast. We extend the study of Scheuerer and Hamill (2015) to some degree and additionally compute the Diebold-Mariano test statistic for each of the 10 repetitions. The resulting test statistic values are also visualized via boxplots.

To evaluate the forecasting distributions we utilize the following multivariate scoring rules:

- Energy score $\mathrm{ES}_{0.1}$, estimated by $\widehat{\mathrm{ES}}_{0.1}^{band}$,
- Energy score $\mathrm{ES}_{0.5}$, estimated by $\widehat{\mathrm{ES}}_{0.5}^{band}$,
- Energy score ES_{1}, estimated by $\widehat{\mathrm{ES}}_{1}^{band}$,
- Energy score $\mathrm{ES}_{1.5}$, estimated by $\widehat{\mathrm{ES}}_{1.5}^{band}$,
- Variogram score $\mathrm{VS}_{W,0.5}$, estimated by $\widehat{\mathrm{VS}}_{W,0.5}$ with W the inverse distance weighting matrix, i.e. $W = (w_{ij})_{ij}$ with $w_{ij} = 1/\sqrt{|i-j|}$,
- Variogram score $\mathrm{VS}_{W,1}$, estimated by $\widehat{\mathrm{VS}}_{W,1}$ with W the inverse distance weighting matrix,
- Variogram score $\mathrm{VS}_{W,2}$, estimated by $\widehat{\mathrm{VS}}_{W,2}$ with W the inverse distance weighting matrix,
- Dawid-Sebastiani score DSS, estimated by $\widehat{\mathrm{DSS}}$.

7.4.1. Miscalibrated marginal distributions

We already noted that the variogram score is unable to detect a bias that is the same for all components, see Subsection 7.3.1. Therefore, we consider a situation where this simple type of bias has been removed.
Specifically, let the observations be realizations of a Gaussian distribution $\mathbf{Y}$ of dimension $d = 5$ with zero mean, unit variance and correlation structure

$$\operatorname{corr}(Y_i, Y_j) = \exp\left(-\frac{|i-j|}{r}\right), \quad i, j = 1, \ldots, d. \tag{7.2}$$

In our simulation study we take $r = 3$. To compare the sensitivity of the different scoring rules to misspecifications of means and variances, we generate forecasts $\mathbf{X}$ with the true exponential correlation structure and

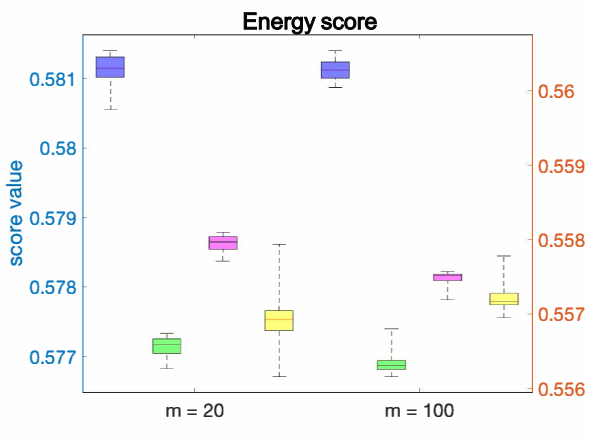
1. correct variances but biased means: $\mu_{\mathbf{X}} = (-0.5, -0.25, 0, 0.25, 0.5)'$,
2. correct means and variances,
3. correct means but too large variances: $\sigma^2_{\mathbf{X}} = 1.5$,
4. correct means but too small variances: $\sigma^2_{\mathbf{X}} = 0.6667$.

The resulting box plots are shown in Figure 7.11. Firstly, we note that the score values improve for all three scoring rules with an increasing ensemble size. This shows that the finite sample representation of the forecasting distribution $F_{\mathbf{X}}$ has a noticable effect on the score value.
However, this sampling effect does not have an impact on qualitive conclusions about the predictive performance of the different forecasts, see Scheuerer and Hamill (2015).
Regarding the Dawid-Sebastiani score, a substantial change of the score values due to the different ensemble size can be observed. Note the different scales for $m = 20$ and $m = 100$, see Figure 7.11h.
The approximation of the process mean $\mu_{\mathbf{X}}$ and covariance structure $\Sigma_{\mathbf{X}}$ by the empirical means and covariances from the small sample is so poor that it leads to wrong conclusions about the predictive accuracy of the forecasts. The score value corresponding to the overdispersive forecast is smaller than the score of the correctly reported forecast, see Figure 7.11h.

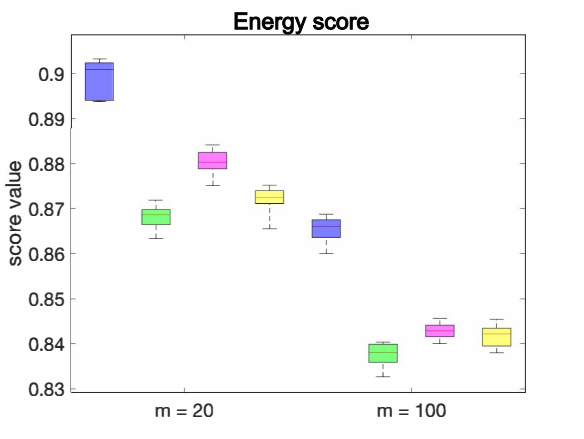
For a larger ensemble size the score bias due to an insufficient representation of the forecasting distribution $F_{\mathbf{X}}$ plays a smaller role and the Dawid-Sebastiani score discriminates well between the calibrated and uncalibrated forecasts.
The energy score effectively detects the erroneous linear trend corresponding to the forecasts simulated according to 1) for all coeffecients β. However, the separation between the calibrated and over-/underdispersive forecasts is less dispersive for a greater parameter β.Thus, the choice of the coefficient β clearly has an influence on the discrimination ability of the energy score in this simulation study. Note that in particular

the boxes corresponding to the over- and underdispersive forecasts separate from the box corresponding to the calibrated forecast for $\beta = 0.1$ very clearly in contrast to the overlapping boxes for $\beta = 1.5$.

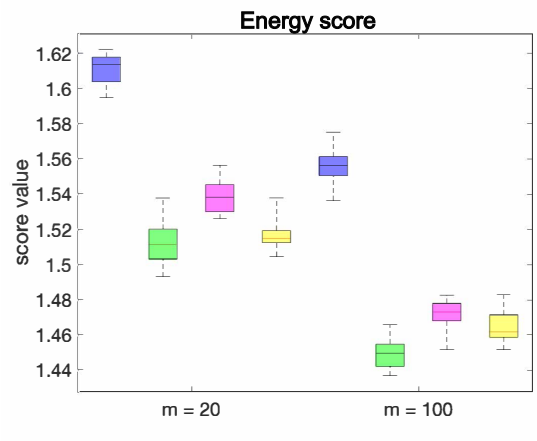
Among the variogram scores of different orders p, the variogram score of order $p = 0.5$ has the best discrimination abilities. It identifies the miscalibrated mean less clearly than the energy score, but it is more effective in detecting over- and underdispersion. The variogram score of order $p = 1$ still has good discrimination abilities and detects all types of miscalibration adequately.
It is noticeable that with an increasing order p the random variations between scores obtained by the identical setup become larger, see Figure 7.11g, and blur the systematic differences between the scores of the calibrated and uncalibrated forecasts.
Recall that the variogram score is not strictly proper. In the present situation, for instance, the effects of an erroneus linear trend and underdispersion cancel out. For $p = 2$ this can directly be seen from (3.22).
Therefore, the authors emphasize that an analysis of the marginal distributions using univariate scoring rules should precede the analysis of multivariate properties, see Scheuerer and Hamill (2015).



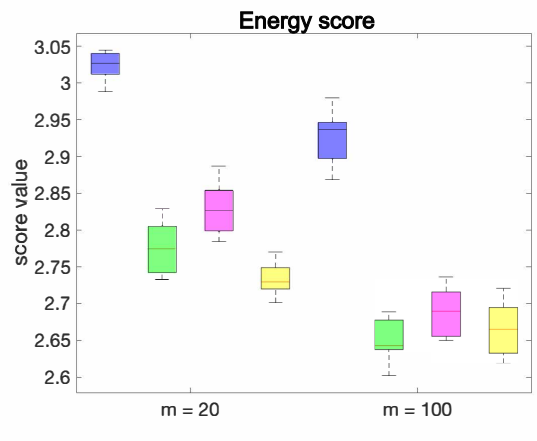
(a) Energy score with $\beta = 0.1$



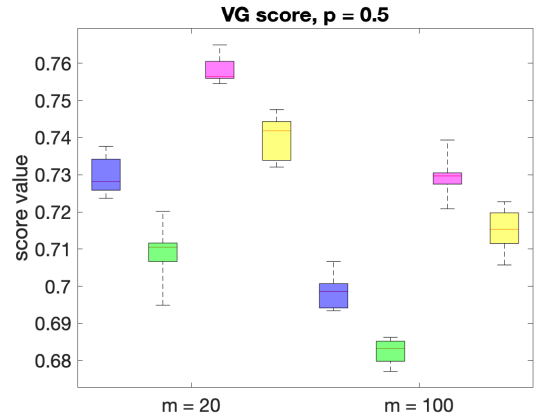
(b) Energy score with $\beta = 0.5$



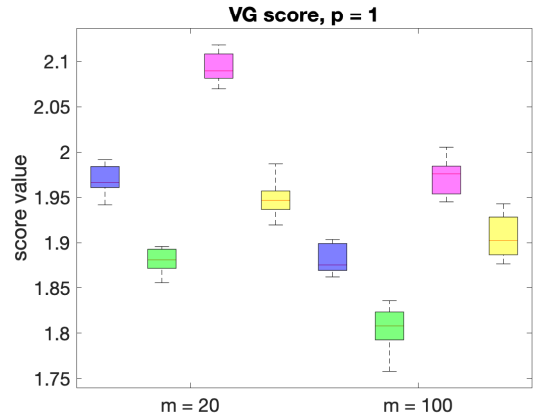
(c) Energy score with $\beta = 1$



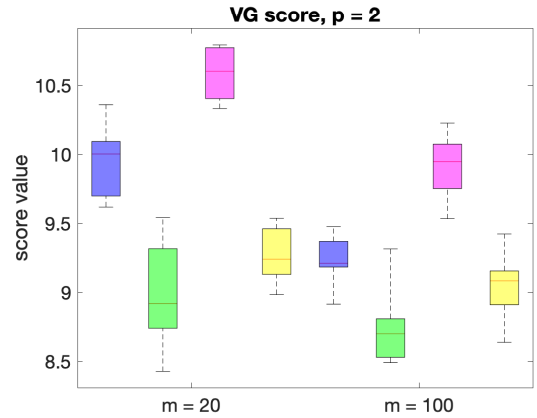
(d) Energy score with $\beta = 1.5$



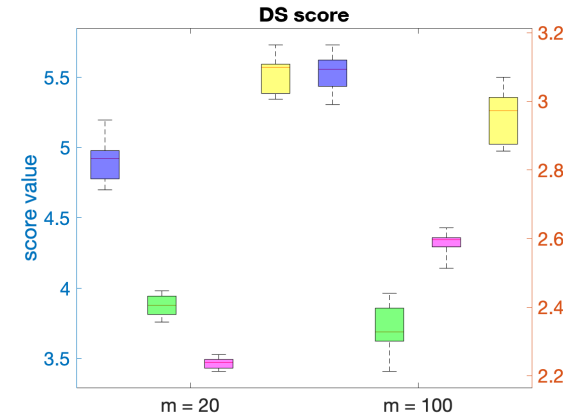
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



(g) VG score of order $p = 2$



(h) DS score

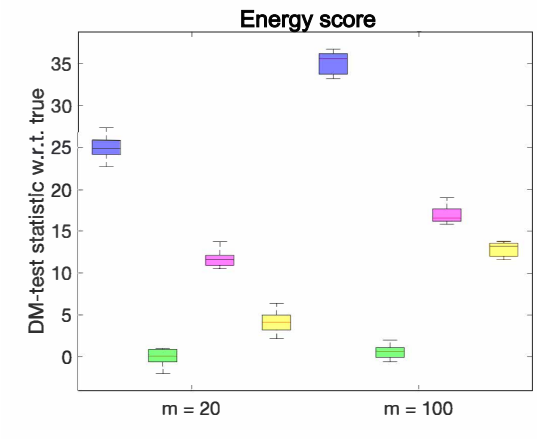
Figure 7.11.: Energy score with coefficient $\beta = 0.5$ (a), $\beta = 1$ (b), and $\beta = 1.5$ (c), the Dawid-Sebastiani score (d), and the variogram score of order $p = 0.5$ (e), $p = 1$ (f), and $p = 2$ (g) for ensemble sizes $m = 20$ and $m = 100$. The boxplots corresponding to the mean-biased, correct, over- and underdispersive forecasts are blue, green, magenta, and yellow, respectively. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes.

After considering the score values of the forecasting experiments, we apply the Diebold-Mariano test to each forecasting experiment. This means we compare the different forecasting distributions with the true underlying process to test if they differ significantly. To this end we compute the values of the DM-test statistic for the different forecasting models 1) - 4) mentioned beforehand.

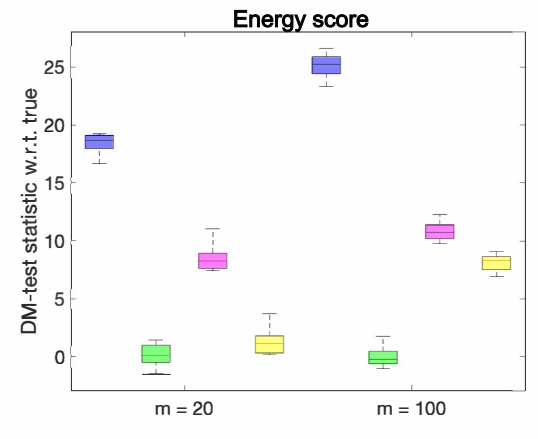
Regarding the energy score, there is an impact on the qualitative conclusions about the predictive performance of the different forecasts due to the sampling size except for the energy score with parameter $\beta = 0.1$. For the smaller forecasting ensemble $m = 20$ the energy score is not able to discriminate the calibrated forecast and the underdispersive forecast which corresponds the previous results. Only for $\beta = 0.1$ all DM-test statistic values are greater than 1.96, that is the energy score with parameter $\beta = 0.1$ is clearly ably to identify the true distribution. The values of the DM-test statistic corresponding to the underdispersive forecast are a little bigger than the values corresponding to the uncalibrated forecast.
The forecast with miscalibrated mean and the overdispersive forecast are detected for all coefficients β for both ensemble sizes. For the large ensemble size $m = 100$ the energy score is able to clearly detect the calibrated forecast for all parameters β. As noted above, the energy score is very effective in detecting an erroneos linear trend and the values of the DM-test statistic corresponding to the forecast with miscalibrated mean are the greatest which holds for all coefficient β of the energy score. However, the values of the DM-test statistic corresponding to the over- and underdispersive forecast obtained by the energy score with coefficient $\beta = 0.1$ are greater than the values obtained by the energy score with coefficient $\beta = 0.5$, $\beta = 1$ and $\beta = 1.5$ in this order. The values obtained by the latter are the smallest and the separation between the calibrated and the underdispersive forecast is rather indistinct.

Also the variogram score is clearly able to detect the calibrated forecast. The values of the DM-test statistic corresponding to the forecast with a biased mean are smaller than the values obtained by the energy score, but the null hypothesis of equal predictive performance can clearly be rejected. It also can be observed that the variogram score is more effective in detecting over- and undispersion than the energy score.
Note that for the larger ensemble size $m = 100$ the variogram score performs similarly for all orders p when the score values are assessed with the Diebold-Mariano test. For the smaller sample size the values of the DM-test statistic corresponding to the underdispersive forecast and obtained by the variogram score of order $p = 0.5$ are greater than for the orders $p = 1$ and $p = 2$. Even for the variogram score of order $p = 2$, which seems to be the least sensitive variogram score, the values of the DM-test statistic corresponding to the underdispersive forecast are greater than 3, i.e. the null hypothesis of equal predictive accuracy with respect to the true distribution is clearly rejected even in this case.
In contrast to the previous statement, which was merely based on the absolute score values, the Dawid-Sebastiani score is able to detect the calibrated forecast for both

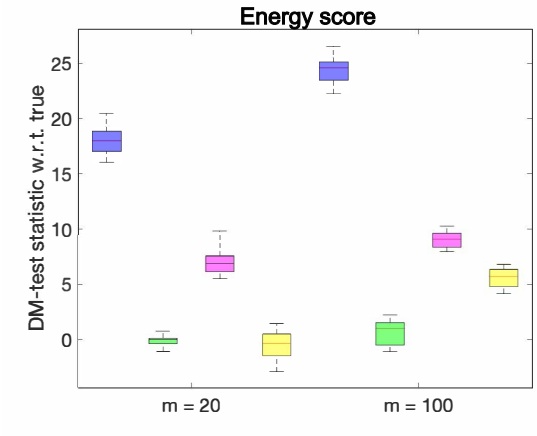
ensemble sizes.
However, note that for $m = 20$ the values of the DM-test statistic for the overdispersive forecast are negative and all smaller than -4.75, i.e. the null hypothesis of equal predictive accuracy compared to the perfect forecast is rejected for all resulting values.



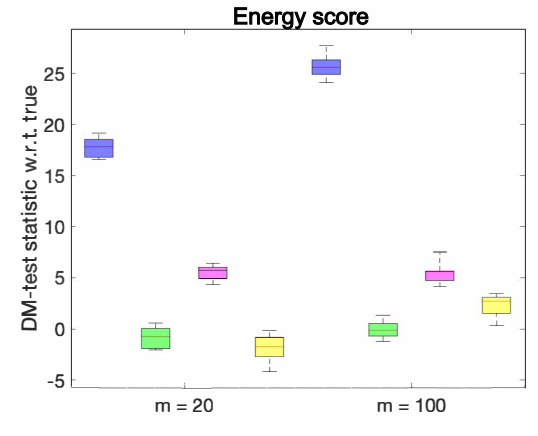
(a) Energy score with $\beta = 0.1$



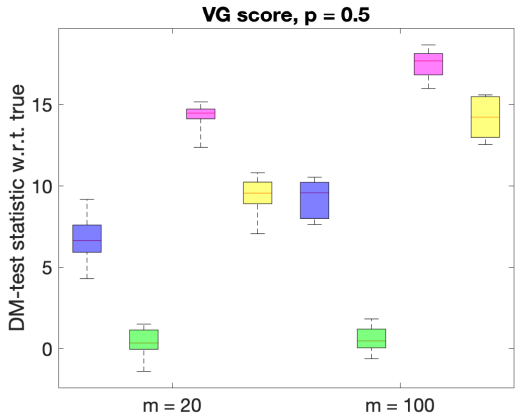
(b) Energy score with $\beta = 0.5$



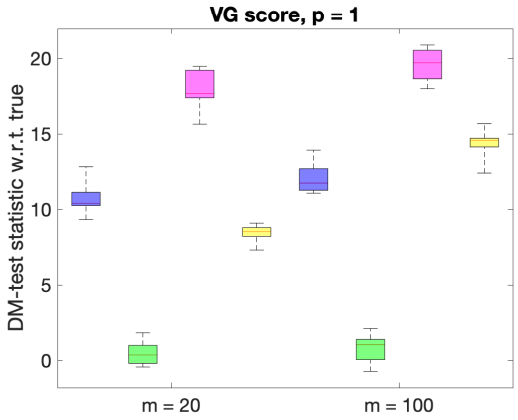
(c) Energy score with $\beta = 1$



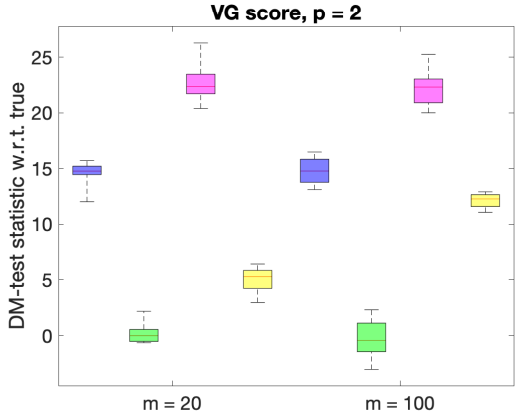
(d) Energy score with $\beta = 1.5$



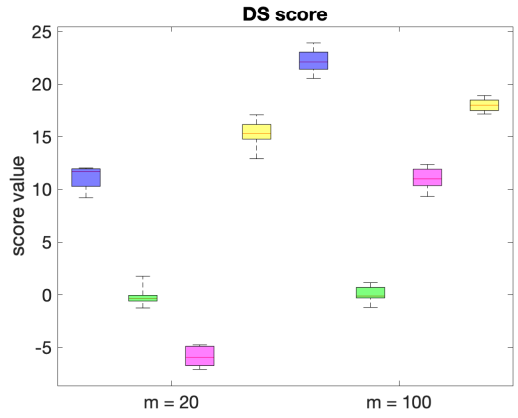
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



(g) VG score of order $p = 2$



(h) DS score

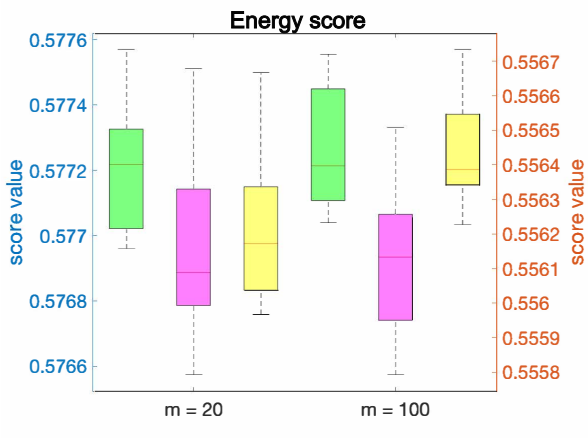
Figure 7.12.: DM-test statistic values calculated with the energy score with coefficient $\beta = 0.5$ (a), $\beta = 1$ (b), and $\beta = 1.5$ (c), the Dawid-Sebastiani score (d), and the variogram score of order $p = 0.5$ (e), $p = 1$ (f), and $p = 2$ (g) for ensemble sizes $m = 20$ and $m = 100$. The boxplots corresponding to the mean-biased, correct, over- and underdispersive forecasts are blue, green, magenta, and yellow, respectively.

7.4.2. Miscalibrated correlation strength

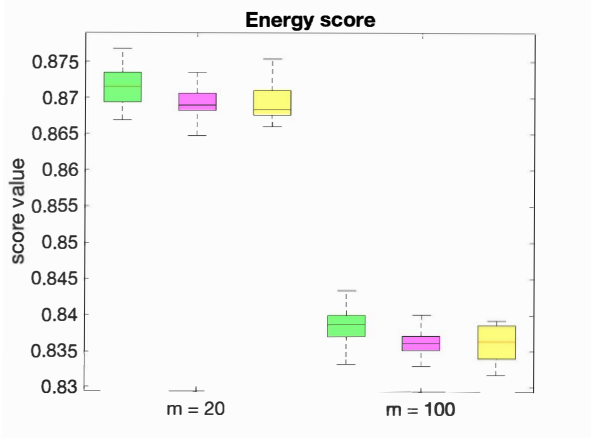
In the next experiment we focus on the correlation structure of the multivariate variable of interest. We study the ability of the different scoring rules to detect whether the correlations between the different components of the forecast are too weak, adequate or too strong with respect to the true distribution of the observation.
Furthermore, we study the influence of increasing the dimension from $d = 5$ to $d = 15$ on the discrimination ability of the different scoring rules.
Again, we consider a zero mean, unit variance AR(1) process with correlation function given in (7.2). As above, we choose $r = 3$ for the true underlying model and compare

ensemble forecasts with the same correlation model, but with $r = 2$, $r = 3$, and $r = 4.5$. This forecasting experiment confirms the conclusion of Pinson and Tastu (2013) that the energy score is not able to discriminate multivariate forecasts that only differ with respect to their correlations between the individual components regardless of the coefficient β, see Figure 7.13b, Figure 7.13c, Figure 7.13b, Figure 7.14b, Figure 7.14c, and Figure 7.14d.
Solely the boxes corresponding to the energy score with parameter $\beta = 0.1$ separate for both dimension sizes for the larger ensemble size $m = 100$. Thus, we can infer that the energy score with $\beta = 0.1$ is able to discriminate between the different forecasting distributions, see Figure 7.13a, and Figure 7.14a.

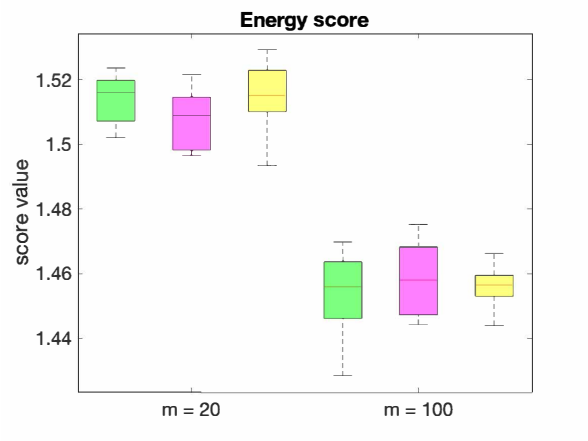
Concerning the Dawid-Sebastiani score, see Figure 7.13h and Figure 7.14h, the conclusion is the same as in the previous experiment. The approximation of the mean $\mu_{\mathbf{Y}}$ and covariance structure $\Sigma_{\mathbf{Y}}$ by the sample mean und sample covariance is rather inaccurate for a small ensemble. Hence, the corresponding score values lead to false conclusions about the predictive performance of the forecasts.
This representation issue is much less significant for the variogram score. For $p = 0.5$ and $p = 1$ the variogram score discriminates well between the calibrated and uncalibrated forecast for both ensemble sizes.
Even the variogram score of order $p = 2$ outperforms the energy score for all parameters except $\beta = 0.1$.



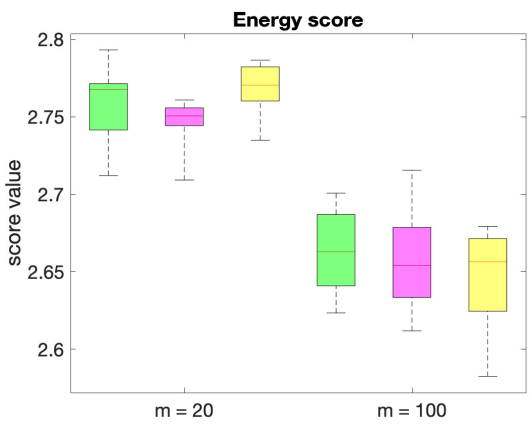
(a) Energy score with $\beta = 0.1$



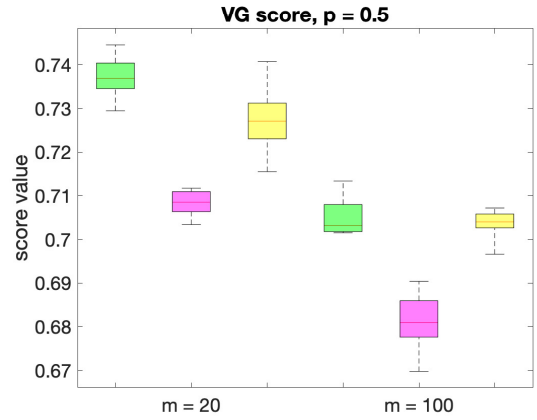
(b) Energy score with $\beta = 0.5$



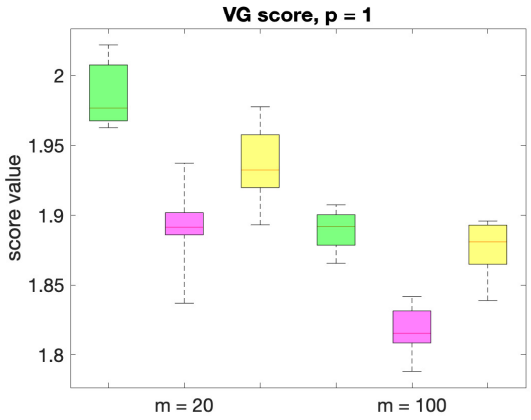
(c) Energy score with $\beta = 1$



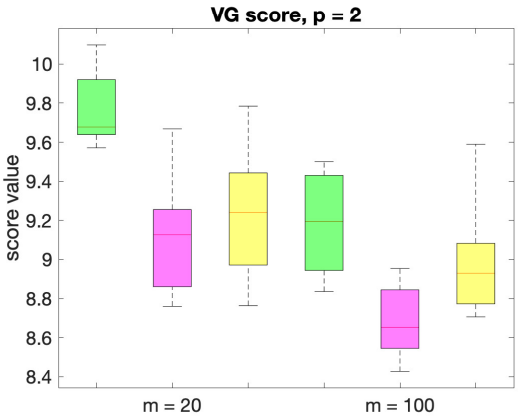
(d) Energy score with $\beta = 1.5$



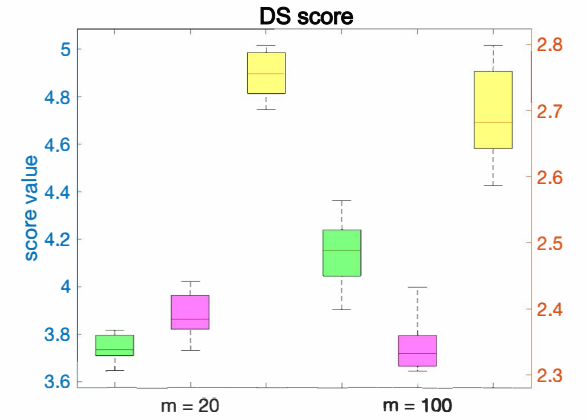
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



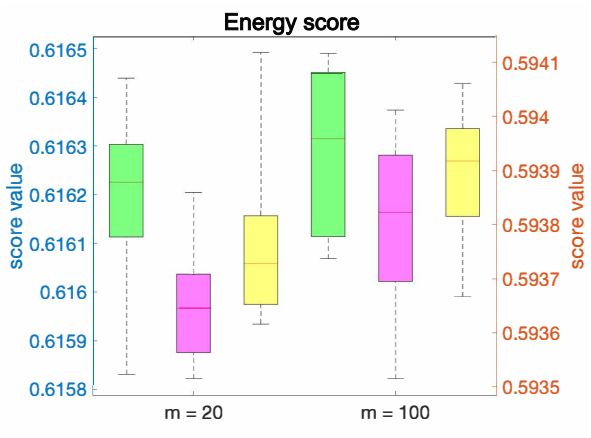
(g) VG score of order $p = 2$



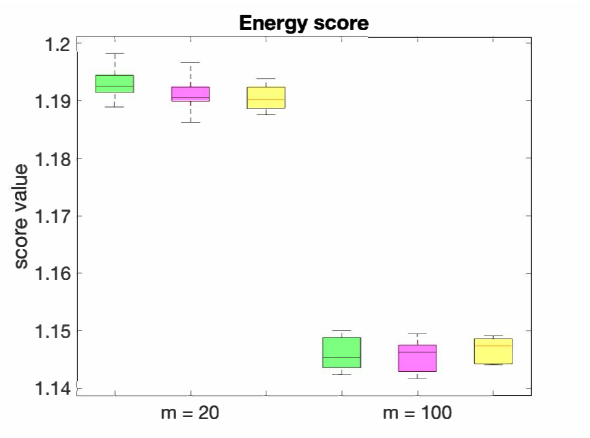
(h) DS score

Figure 7.13.: As in Figure 7.11, but for forecasts with too weak (green), accurate (magenta), and too strong (yellow) correlations compared to the observations for dimension $d = 5$.

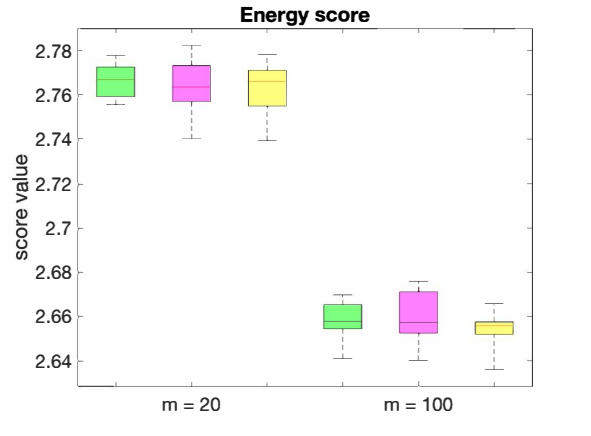
Note that a larger dimension size d, see Figure 7.14, has a slightly negative effect on the discrimination ability of the variogram score. At first this seems surprising since a larger dimension size yields more data that is used for the calculation of the variogram score. However, increasing the number of summands in (5.1) does not lead to an averaging of sampling error as our definition of $\mathrm{VS}_{W,p}$ does not make any assumptions about the correlation structure of forecasts and observations.



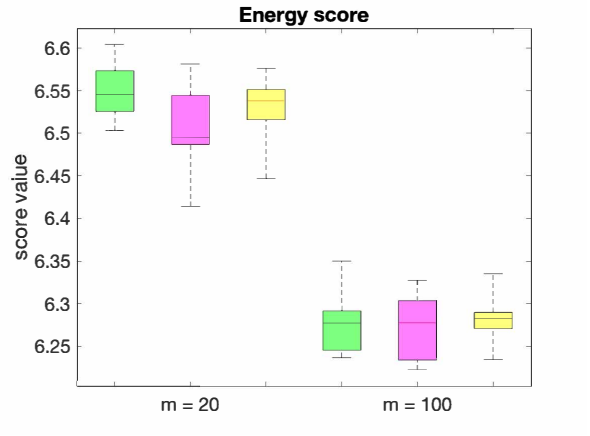
(a) Energy score with $\beta = 0.1$



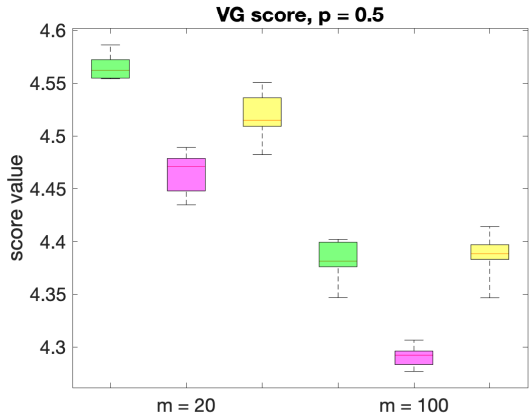
(b) Energy score with $\beta = 0.5$



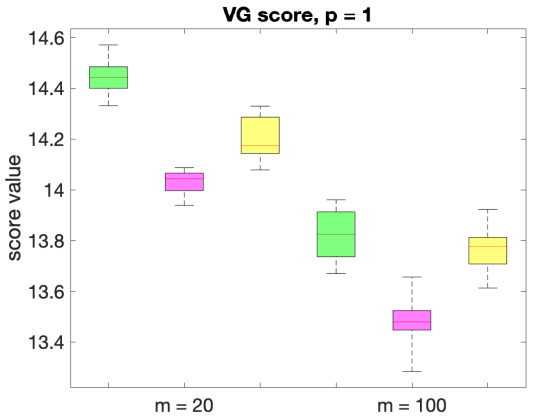
(c) Energy score with $\beta = 1$



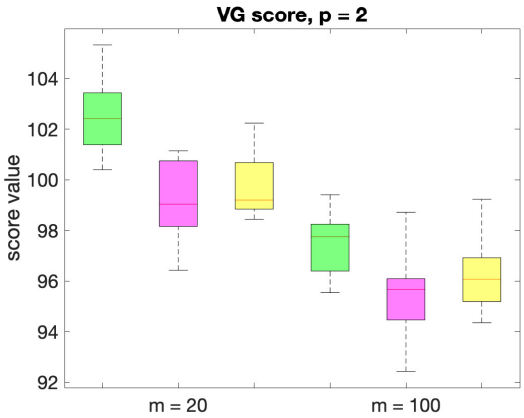
(d) Energy score with $\beta = 1.5$



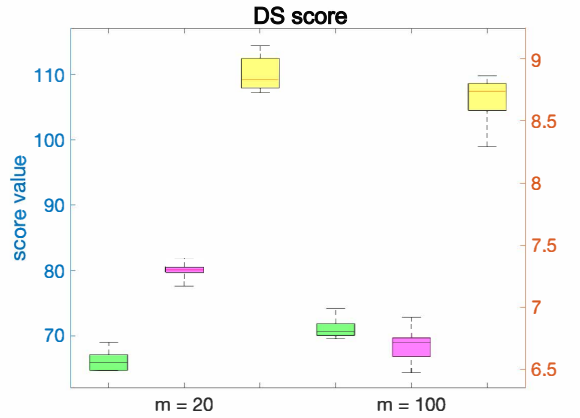
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



(g) VG score of order $p = 2$



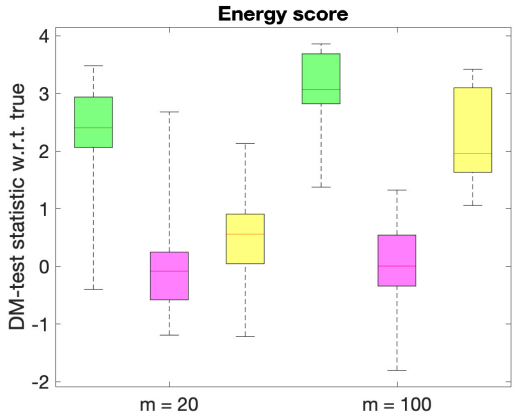
(h) DS score

Figure 7.14.: As in Figure 7.13, but for dimension $d = 15$.

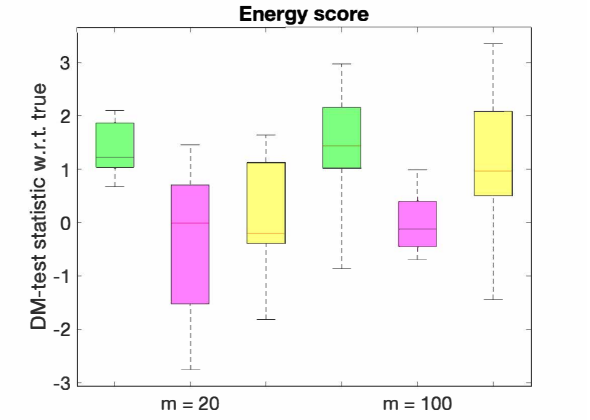
The energy score with $\beta = 0.1$ is clearly able to discriminate between the forecasts with correct and incorrect specified correlation when the Diebold-Mariano test is applied, see Figure 7.15a, and even the energy score with $\beta = 0.5$ to some degree, see Figure 7.15b.
For the energy score with $\beta = 0.1$ the majority of DM-test statistic values are outside the range from -1.96 to 1.96, so mostly the null hypothesis of equal predictive accuracy in comparison to the calibrated forecast is correctly rejected. However, this is not true for the energy score with $\beta = 0.5$, $\beta = 1$ and $\beta = 1.5$.

The variogram score is able to identify the calibrated forecast for all orders p. However, note that for a larger order p the values of the DM-test statistic corresponding to the forecast with too strong correlation are smaller, see Figure 7.15e, Figure 7.15f and Figure 7.15g. This is particularly true for the smaller ensemble size and also corresponds to our findings above.
As noted above, the score values obtained by the Dawid-Sebastiani score for the small

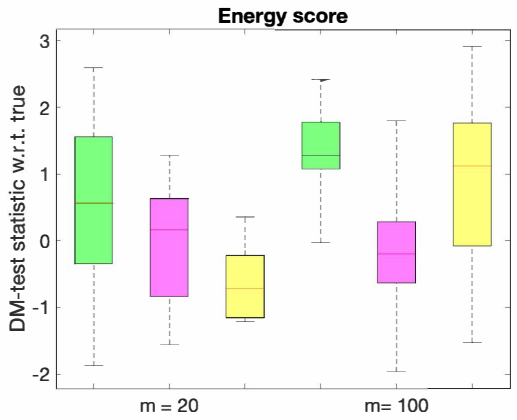
ensemble size are the smallest for the forecast with too weak correlation. However, most of the values of the corresponding DM-test statistic are smaller than -1.96, i.e. in most cases the null-hypothesis of equal predictive accuracy in comparison with the perfect forecast is correctly rejected at significance level $\alpha = 5\,\%$.
It is also noticeable that the DM-test statistic values calculated with the Dawid-Sebastiani score take greater values for the forecast with too strong correlation for both ensemble sizes.
Note that the discrimination of the energy score is approximately the same when the dimension size is $d = 15$. Futhermore, the values corresponding to the Dawid-Sebastiani score for the larger dimension $d = 15$ are similar to the values obtained when the dimension size is $d = 5$.



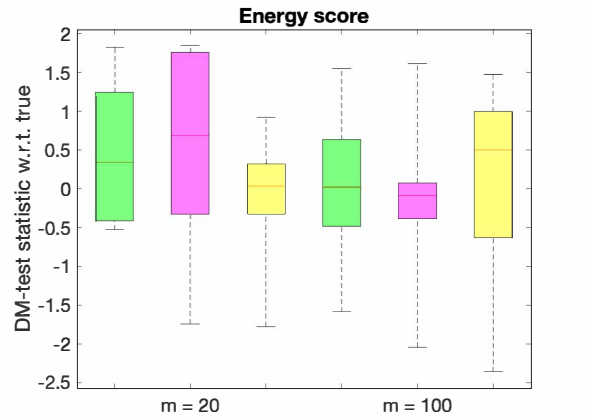
(a) Energy score with $\beta = 0.1$



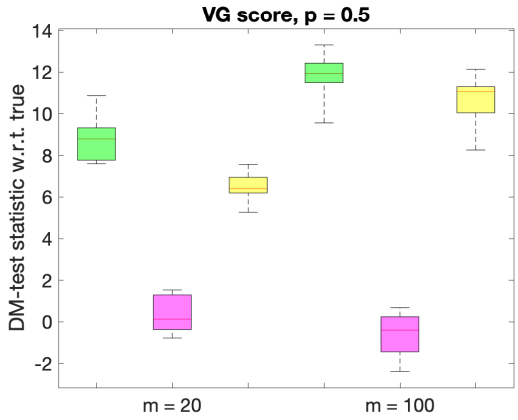
(b) Energy score with $\beta = 0.5$



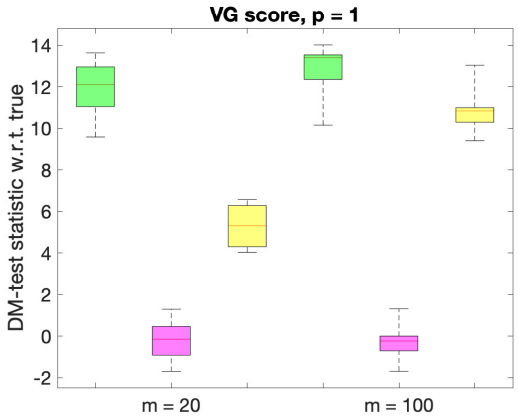
(c) Energy score with $\beta = 1$



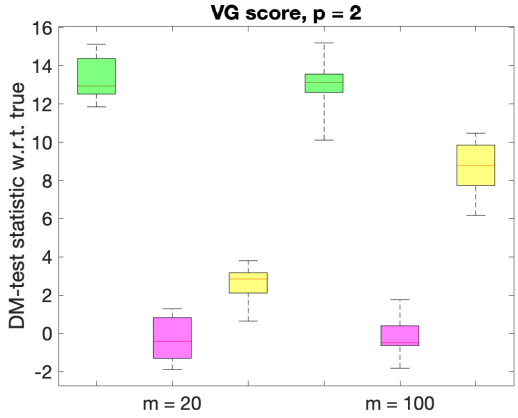
(d) Energy score with $\beta = 1.5$



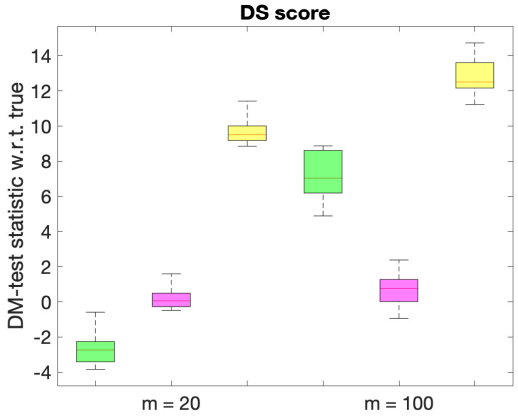
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$

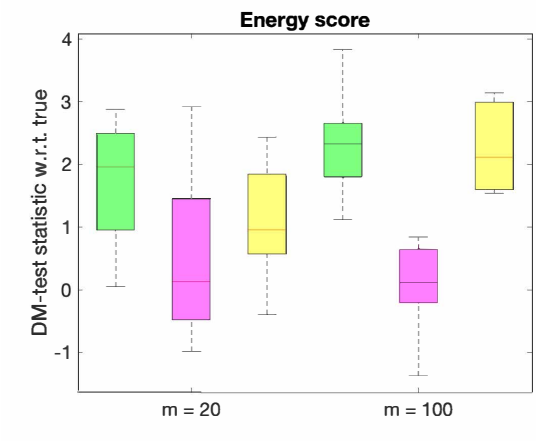


(g) VG score of order $p = 2$

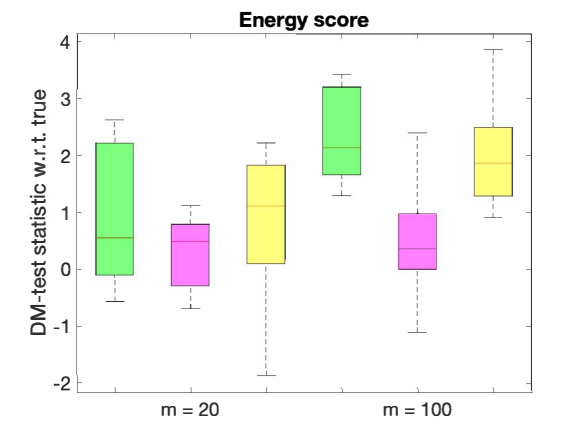


(h) DS score

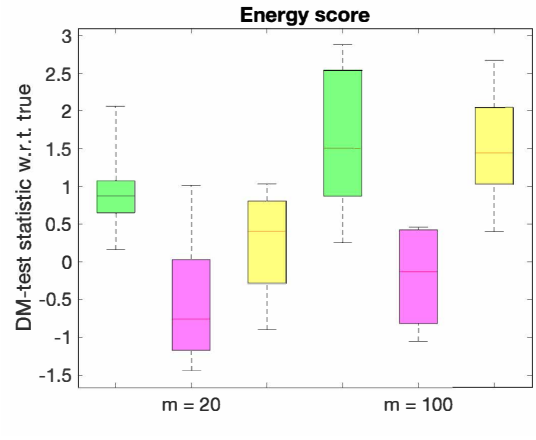
Figure 7.15.: DM-test statistic values for forecasts with too weak (green), accurate (magenta), and too strong (yellow) correlations compared to the observations for dimension $d = 5$.



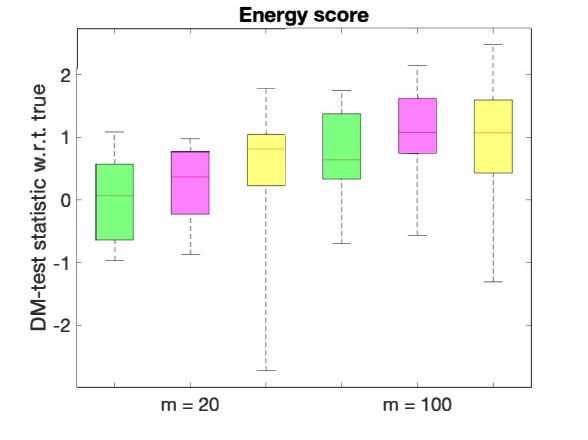
(a) Energy score with $\beta = 0.1$



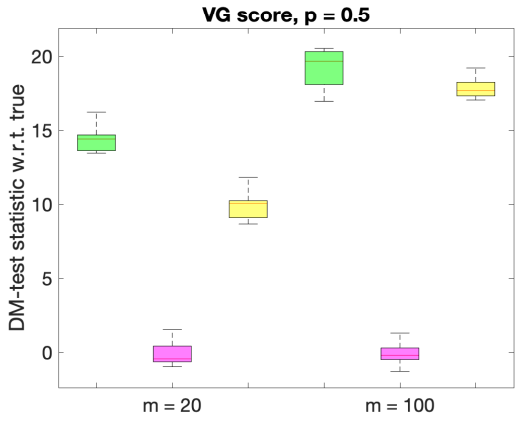
(b) Energy score with $\beta = 0.5$



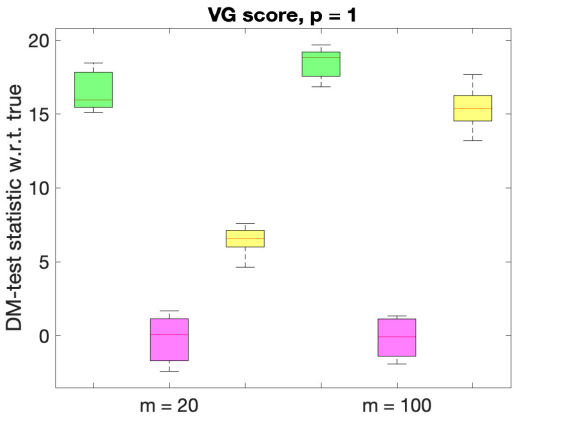
(c) Energy score with $\beta = 1$



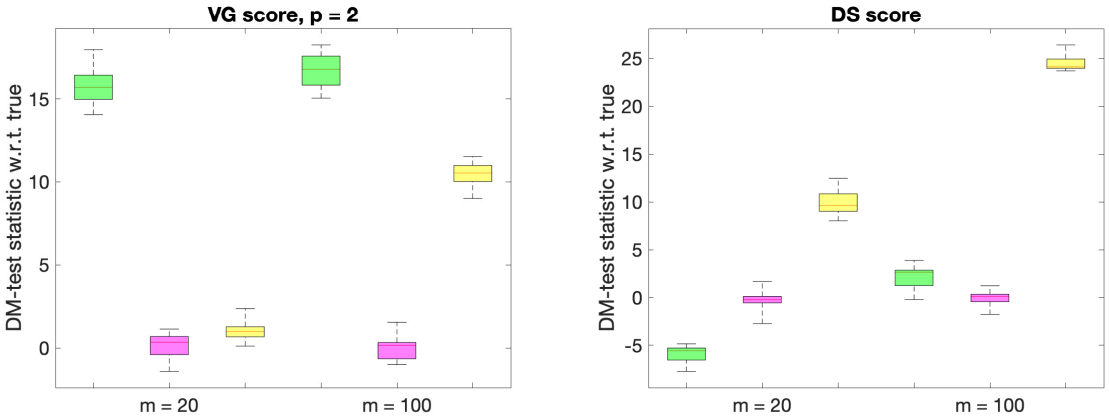
(d) Energy score with $\beta = 1.5$



(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



(g) VG score of order $p = 2$ (h) DS score

Figure 7.16.: Same as Figure 7.15, but for dimension $d = 15$.

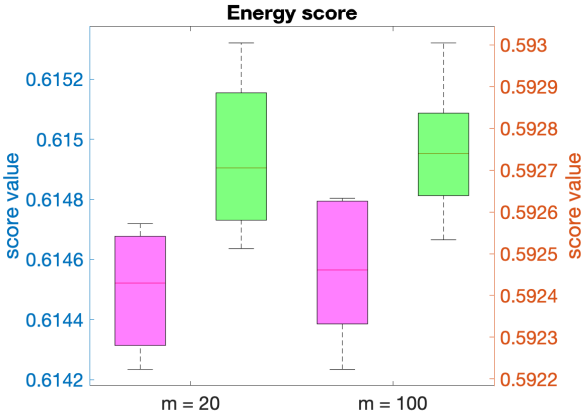
7.4.3. Misspecified correlation model

In this subsection we focus on a misspecified correlation model. Here we not just vary the correlation strength but the entire correlation model. We consider observations with zero mean, unit variance, and correlation function:

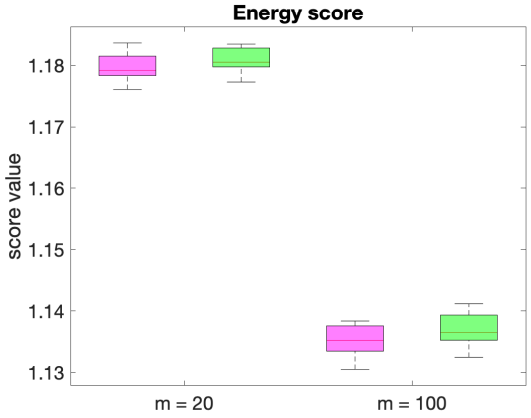
(i) $\operatorname{corr}(Y_i, Y_j) = \left(1 + \frac{|i-j|}{3}\right)^{-1}$, and

(ii) $\operatorname{corr}(Y_i, Y_j) = \exp\left(-\frac{|i-j|}{4}\right)\left[0.75 + 0.25 \cdot \cos\left(\frac{|i-j|\pi}{2}\right)\right]$.

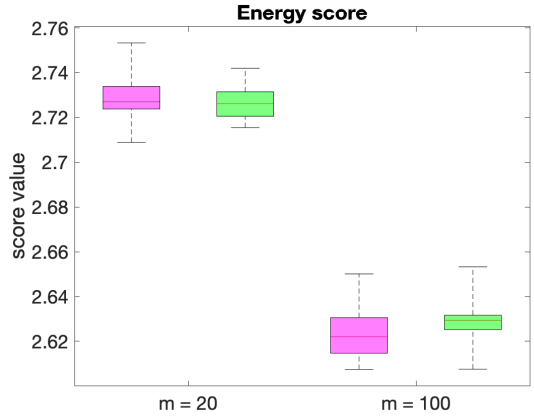
Both forecasting models yield correlations at lag 1 that are quite similar to the true underlying distribution with the exponential model in (7.2) with $r = 3$. However, model (i) has much stronger correlations at larger lags and model (ii) has a periodic component that makes it oscillate around the exponential model.
In the following, we only consider the case where the dimension size is $d = 15$. This experiment confirms many conclusions about the discrimination abilities of the different scoring rules from the preceding forecasting experiments.



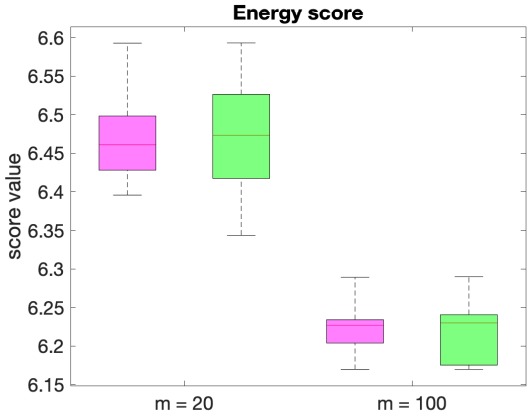
(a) Energy score with $\beta = 0.1$



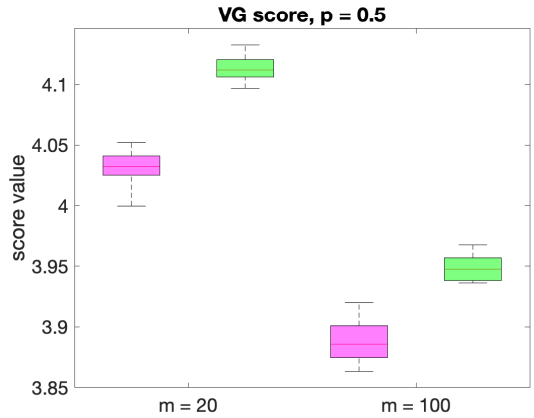
(b) Energy score with $\beta = 0.5$



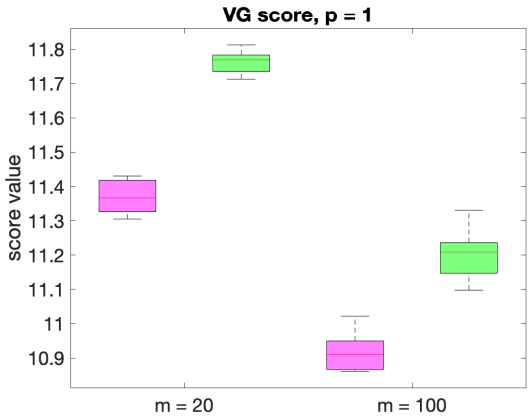
(c) Energy score with $\beta = 1$



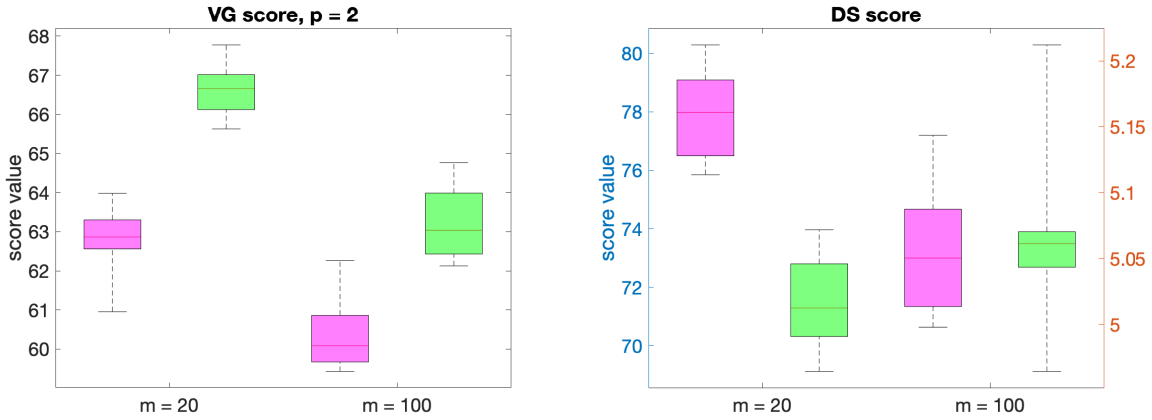
(d) Energy score with $\beta = 1.5$



(e) VG score of order $p = 0.5$



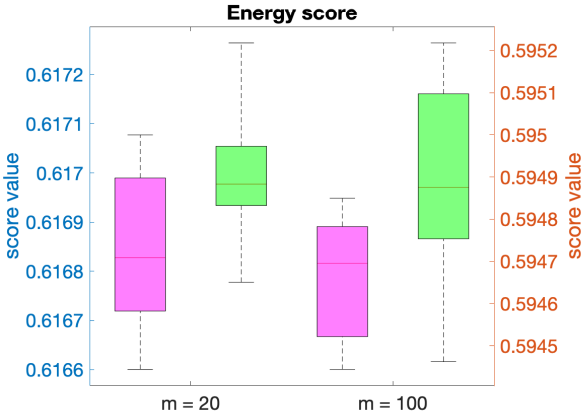
(f) VG score of order $p = 1$



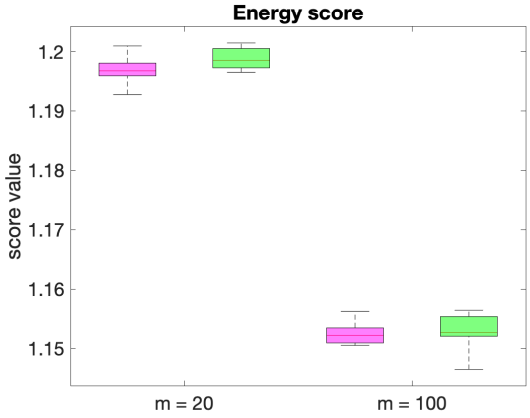
(g) VG score of order $p = 2$ (h) DS score

Figure 7.17.: As in Figure 7.11, but for forecasts with correct (left magenta boxplots), and incorrect (right green boxplots) correlation structure, where the correct correlation function is that using model (i) in Section 7.4.3, and the incorrect correlation function is the exponential model in (7.2) with $r = 3$.

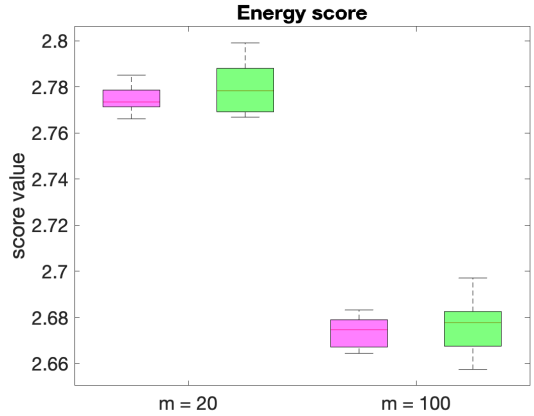
As in the preceding forecasting experiment, the energy score lacks sensitivity to misspecifications of the correlation structure, see Figure 7.17b, Figure 7.17b, Figure 7.17b, Figure 7.18b, Figure 7.18c, and Figure 7.17d. Only the energy score with $\beta = 0.1$ is able to discriminate between the different distributions as the boxes corresponding to the correct and incorrect forecasts separate for ensemble size $m = 100$ to some degree, see Figure 7.17a, and Figure 7.18a. However, this separation is not as clear as for the variogram score or the Dawid-Sebastiani score.
The variogram score distinguishes well between the correct and incorrect specified distributions. Again, the discrimination ability depends on p, where smaller values yield better results.
The Dawid-Sebastiani score has similar issues as in the preceding experiments. When the observations have long-range dependence both ensemble sizes are not sufficient to yield a proper ranking between calibrated and uncalibrated forecast, see Figure 7.17h. Even increasing the ensemble size to $m = 100$ does not reduce the score's representation bias enough to yield a proper ranking between the correct and incorrect forecast.
In the example of the oscillating correlation model on the other hand the Dawid-Sebastiani score separates the two forecasts very well and yields the correct ranking for both ensemble sizes, see Figure 7.18h.



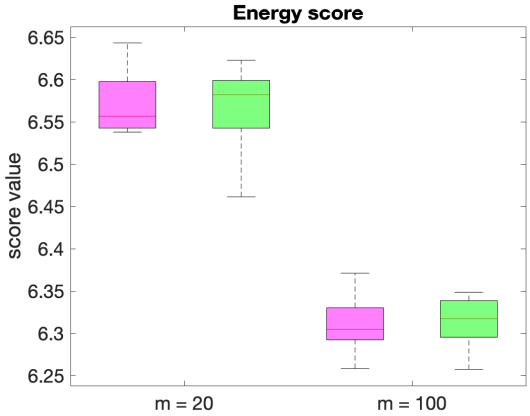
(a) Energy score with $\beta = 0.1$



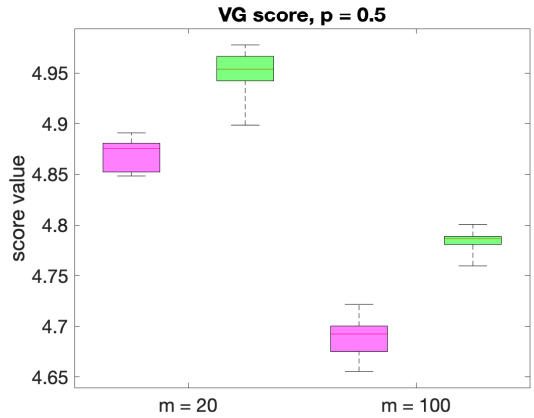
(b) Energy score with $\beta = 0.5$



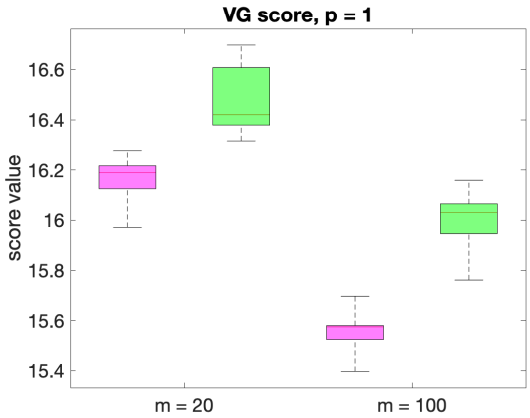
(c) Energy score with $\beta = 1$



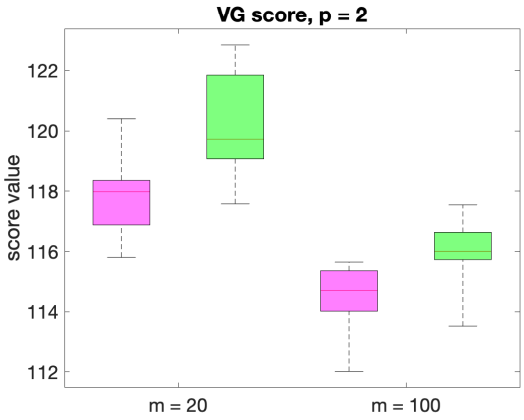
(d) Energy score with $\beta = 1.5$



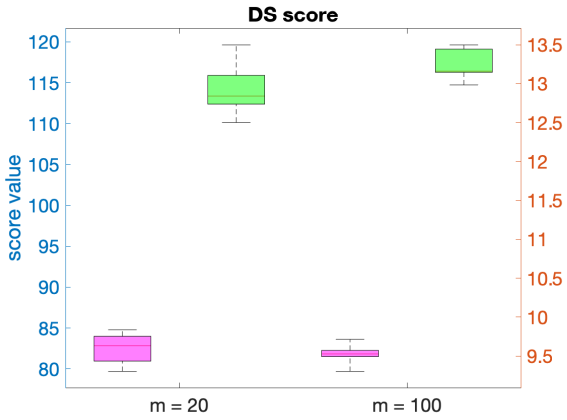
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$

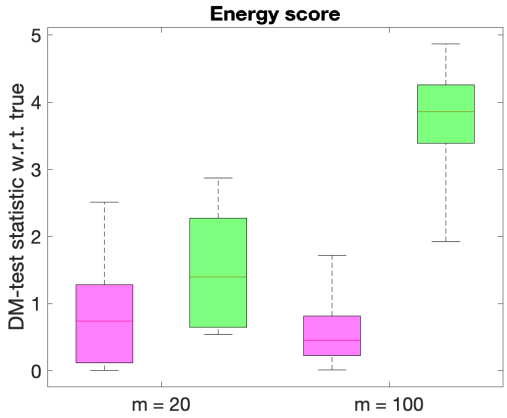


(g) VG score of order $p = 2$

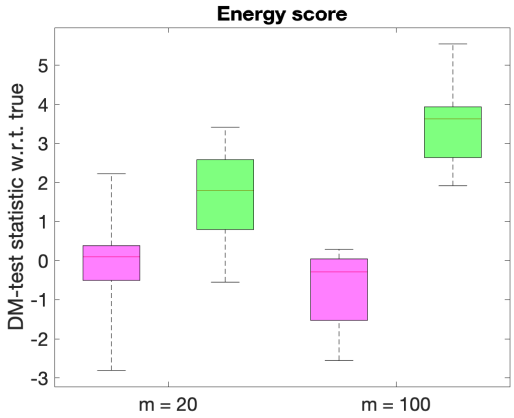


(h) DS score

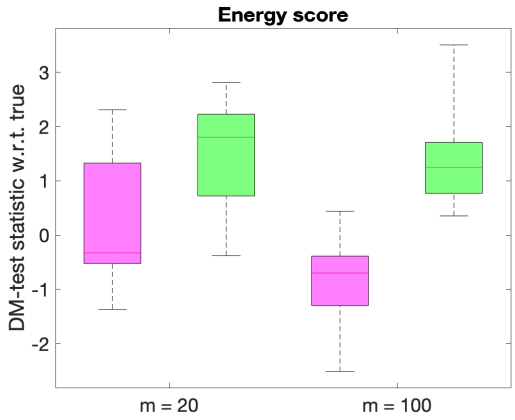
Figure 7.18.: As in Figure 7.11, but for forecasts with correct (left magenta boxplots), and incorrect (right green boxplots) correlation structure, where the correct correlation function is that using model (ii) in Section 7.4.3, and the incorrect correlation function is the exponential model in (7.2) with $r = 3$.



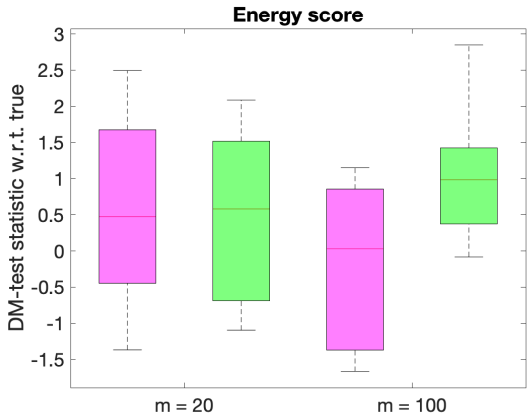
(a) Energy score with $\beta = 0.1$



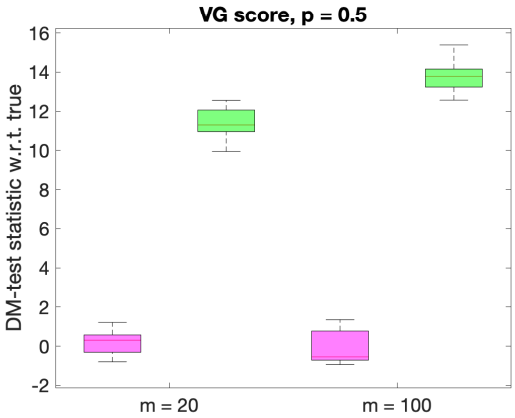
(b) Energy score with $\beta = 0.5$



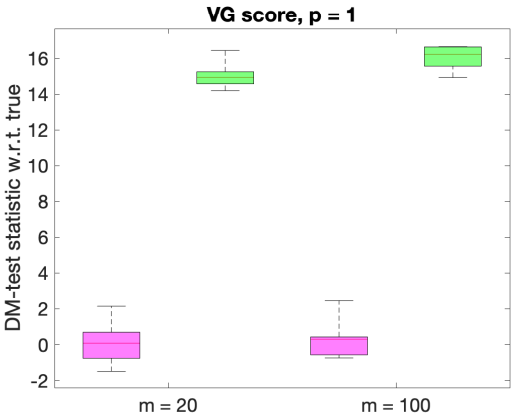
(c) Energy score with $\beta = 1$



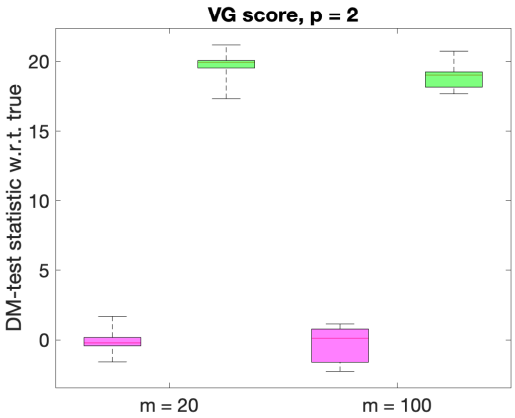
(d) Energy score with $\beta = 1.5$



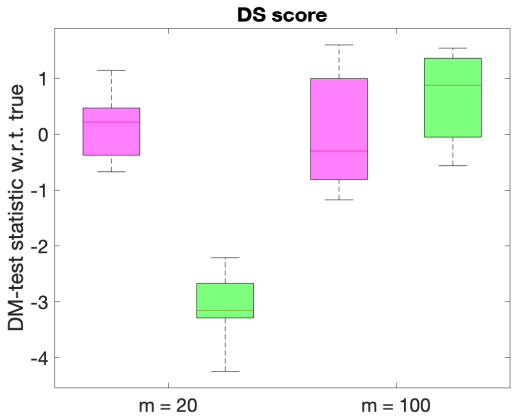
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



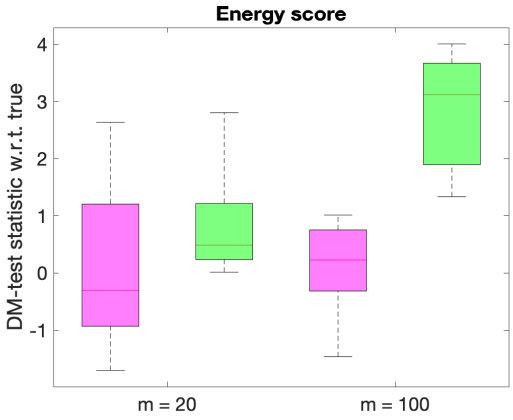
(g) VG score of order $p = 2$



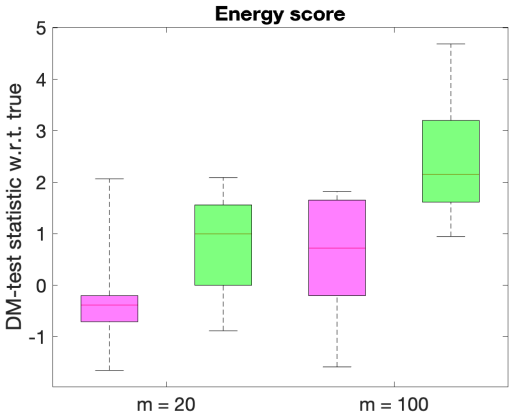
(h) DS score

Figure 7.19.: DM-test statistic values for forecasts with correct (left magenta boxplots), and incorrect (right green boxplots) correlation structure, where the correct correlation function is that using model (i) in Section 7.4.3, and the incorrect correlation function is the exponential model in (7.2) with $r = 3$.

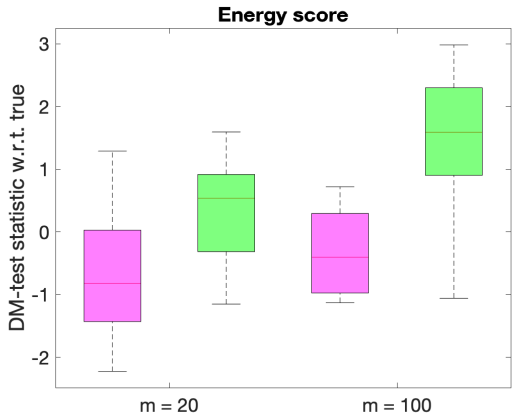
The Diebold-Mariano test confirms the preceding statements. Considering the DM-test statistic the energy score with $\beta = 0.1$ is clearly able to discriminate between the correct and incorrect specified forecasting distributions. Also the energy score with $\beta = 0.5$ is able to discriminate between these distributions to some degree. For $\beta = 0.1$ the DM-test statistic values corresponding to the forecast with the misspecified correlation model are all greater than 1.96 for ensemble size $m = 100$ for both forecasting experiments, see Figure 7.19a, and Figure 7.20a. For $\beta = 0.5$ and ensemble size $m = 100$ in the first forecasting experiment all DM-test statistic values corresponding to the miscalibrated forecast are greater than 1.96 and in the second forecasting experiment the majority of DM-test statistic values. For $\beta = 1$ and $\beta = 1.5$ most DM-test statistic values corresponding to miscalibrated forecast are in the range from -1.96 to 1.96.
So as in the preceding experiment the discrimination ability of energy score depends on β and the energy score with the smallest parameter yields the best discrimination ability.
The DM-test statistic values of the uncalibrated forecast calculated with the variogram score are significantly larger for both ensemble sizes and all orders p similar to the preceding forecasting experiment.
In the example of the correlation model with long-range dependence the Dawid-Sebastiani score is able to identify the correct and incorrect forecast for $m = 20$, see Figure 7.19h, but it may be that in this example the Dawid-Sebastiani score favors the correct forecast by chance due to the finite representation of the predective distribution. This assumption is supported by the fact that for $m = 100$ the Dawid-Sebastiani score is not able to discriminate between the different forecast.



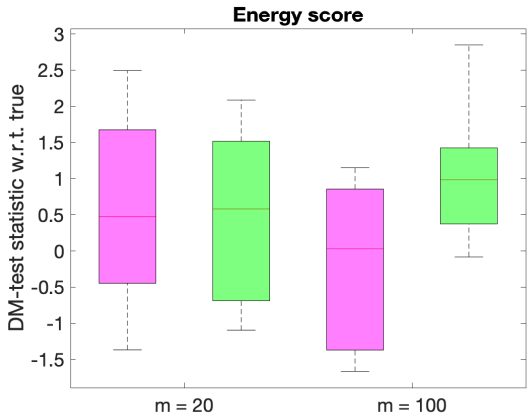
(a) Energy score with $\beta = 0.1$



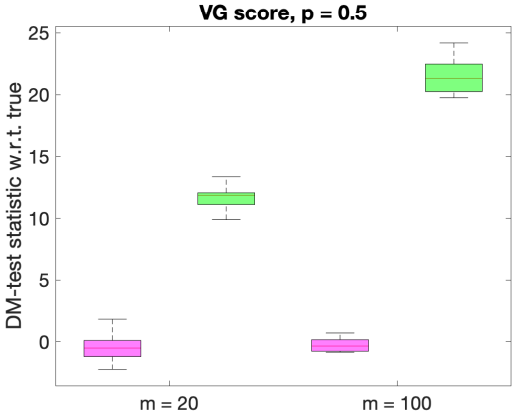
(b) Energy score with $\beta = 0.5$



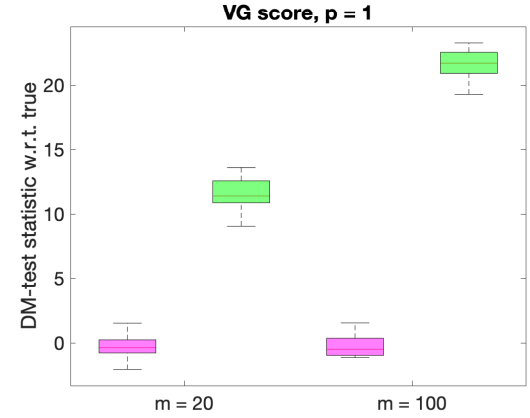
(c) Energy score with $\beta = 1$



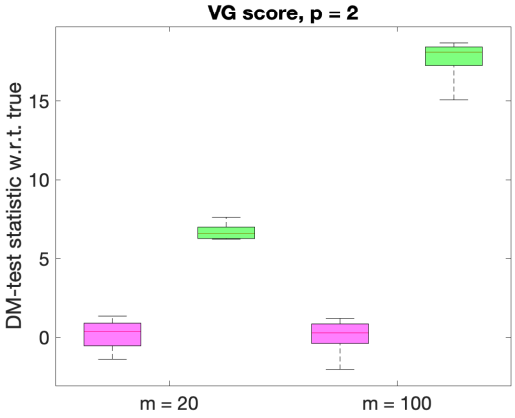
(d) Energy score with $\beta = 1.5$



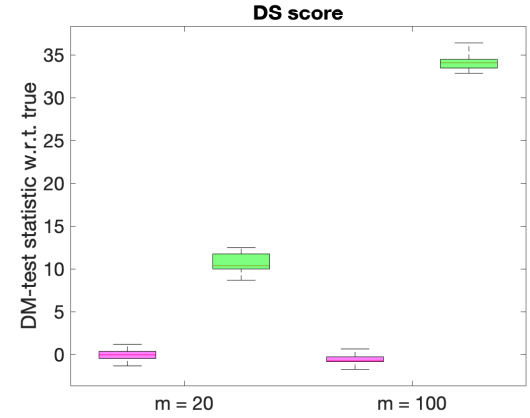
(e) VG score of order $p = 0.5$



(f) VG score of order $p = 1$



(g) VG score of order $p = 2$



(h) DS score

Figure 7.20.: DM-test statistic values for forecasts with correct (left magenta boxplots), and incorrect (right green boxplots) correlation structure, where the correct correlation function is that using model (ii) in Section 7.4.3, and the incorrect correlation function is the exponential model in (7.2) with $r = 3$.

7.4.4. Discussion of the study results

Among the energy scores with different parameters β the energy score with $\beta = 0.1$ clearly has the best discrimination ability. Considering the values of the DM-test statistic this scoring rule is able to discriminate between the true and the misspecified model in all three settings. Thus, we conclude that the standard choice of $\beta = 1$ in application has to be reconsidered as the energy score with a smaller parameter β clearly has a better discrimination ability.
For this conclusion it is not even sufficient to consider the values of the DM-test statistic, even the boxes corresponding to the score values discriminate well.
In this simulation study different orders p of variogram scores were considered. The best results are obtained by $p = 0.5$, while $p = 2$ was clearly not an optimal choice. In Scheuerer and Hamill (2015), the authors state that for Gaussian distributions the choice $p = 0.5$ is optimal since the distribution of $|X_i - X_j|^{0.5}$ is almost perfectly symmetric, and, therefore, has much better sampling properties than the strongly skewed distribution that comes with $p = 2$. If the forecasting distribution itself is already skewed, it might be optimal to choose a smaller power p to obtain a near-symmetric distribution of $|X_i - X_j|^p$. From a qualitative perspective, the energy score with $\beta = 0.1$ has a discrimination ability similar to the variogram score of order $p = 0.5$ and the Dawid-Sebastiani score. From a quantitative perpective, the DM-test statistic values corresponding to the miscalibrated forecast are greater for the variogram score of order $p = 0.5$ and the Dawid-Sebastiani score. However, in all cases the DM-test statistic values obtained by the energy score with $\beta = 0.1$ are sufficiently large to correctly reject the null hypothesis of equal predictive accuracy.
Furthermore, the energy score is the only scoring rule among these three that generally is strictly proper, so we infer that it should be the standard choice to assess probabilistic forecasts.

7.5. Simulation study III: bivariate Gumbel copula

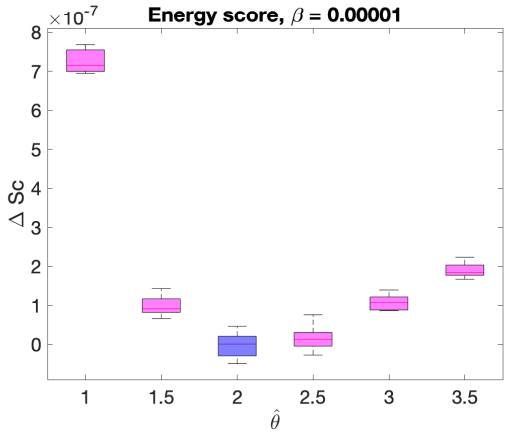
In the previous sections we considered the discrimination ability with respect to Gaussian distributions which are by definition elliptical distributions with a linear dependence structure.
Now we are interested in forecasting a bivariate non-elliptical distribution $\mathbf{Y}$, where the dependence structure of the two components is given by the Gumbel copula

$$C_\theta^{GU}(u, v) = \exp\left(- \left[(-\log u)^\theta + (-\log v)^\theta\right]^{1/\theta}\right),$$

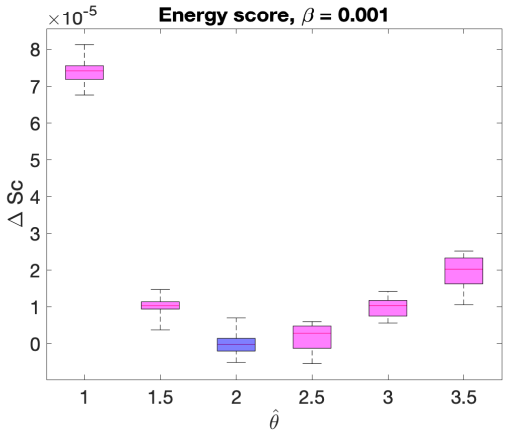
where the parameter $1 < \theta < \infty$ controls the dependence. For a short introduction to copulas we refer to Appendix A.1.

7.5.1. Uniformly distributed marginals

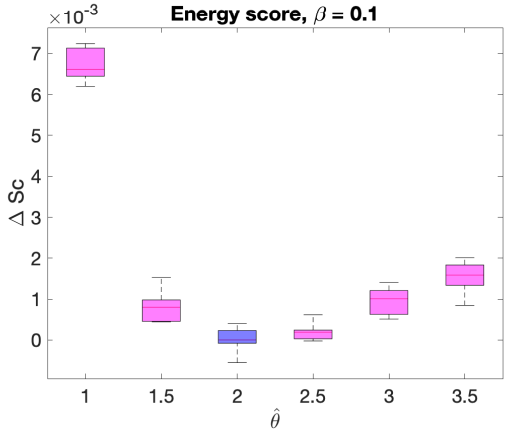
First, we consider a bivariate distribution $\mathbf{Y}$ with uniformly distributed marginals, i.e. we only study the discrimination ability of the energy score, the variogram score and the Dawid-Sebastiani score with respect to the copula. In the following, we assume the distribution of $\mathbf{Y}$ is given by the bivariate Gumbel copula with parameter $\theta = 2$. We suppose the marginals are reported correctly, that is the probabilistic forecasts only differ in the parameter $\hat{\theta}$ which is chosen out of the grid $\hat{\theta} \in \{1, 1.5, \ldots, 3.5\}$. As in the preceding sections, the discrimination ability of the different scoring rules is assessed with the relative change in score value and the Diebold-Mariano test.



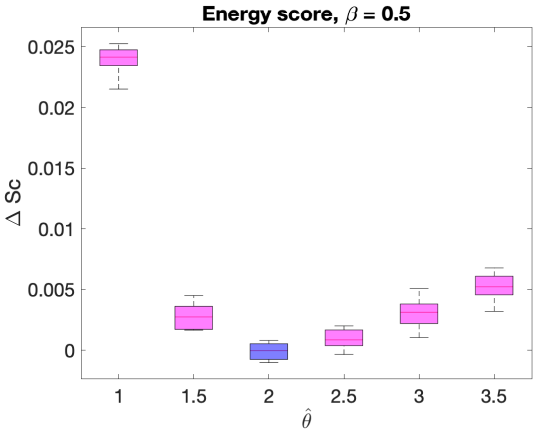
(a) Energy score with $\beta = 0.00001$



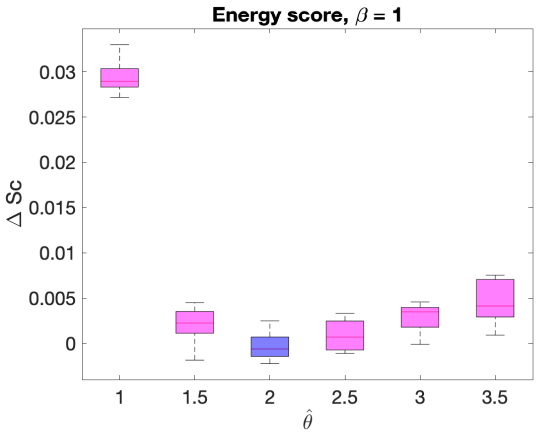
(b) Energy score with $\beta = 0.001$



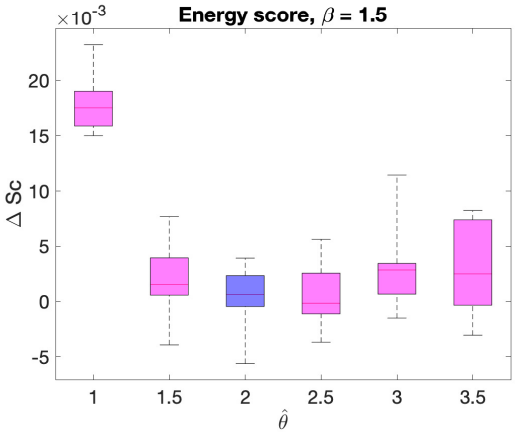
(c) Energy score with $\beta = 0.1$



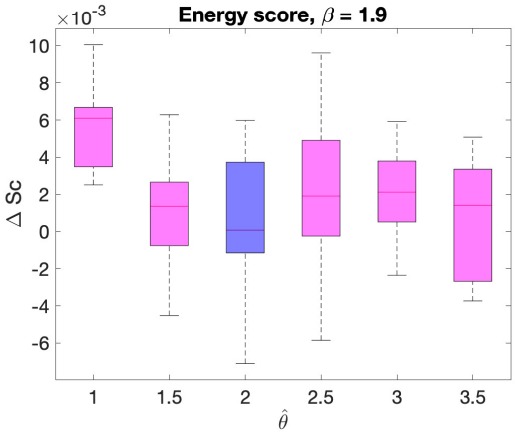
(d) Energy score with $\beta = 0.5$



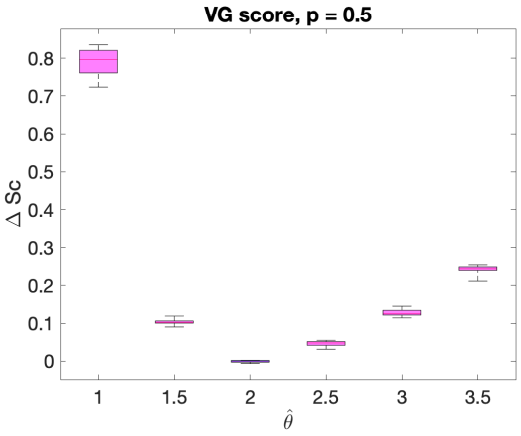
(e) Energy score with $\beta = 1$



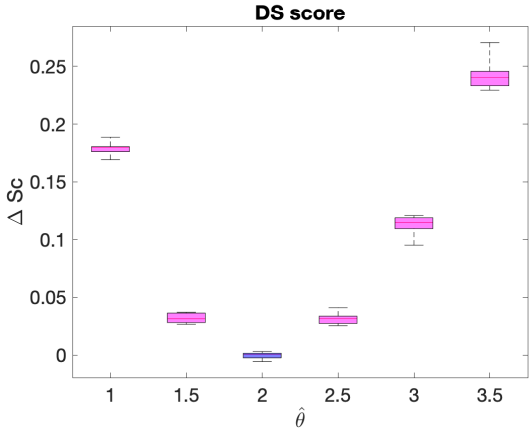
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$



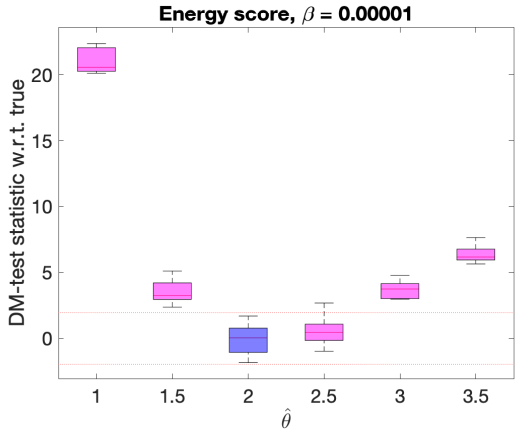
(h) VG score with $p = 0.5$



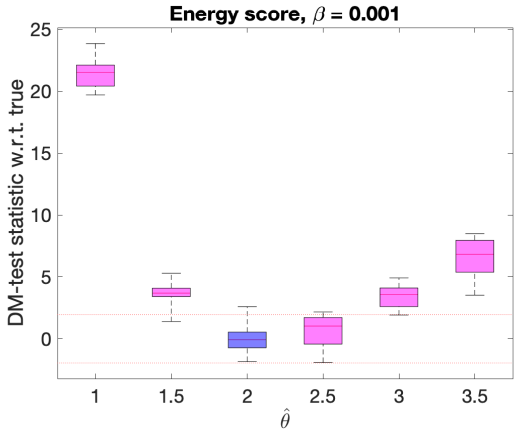
(i) DS score

Figure 7.21.: Energy score with coefficients $\beta = 0.00001$, $\beta = 0.001$, $\beta = 0.1$, $\beta = 0.5$, $\beta = 1$, $\beta = 1.5$, and $\beta = 1.9$ as well as the variogram score with $p = 0.5$, and the Dawid-Sebastiani score assessed with the relative change in score value for an ensemble size $m = 100$. The boxplots corresponding to the forecasts with the correct and incorrect specified parameter $\hat{\theta}$ are blue and magenta, respectively. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes. The true underlying distributions is given by the bivariate Gumbel copula with parameter $\theta = 2$.

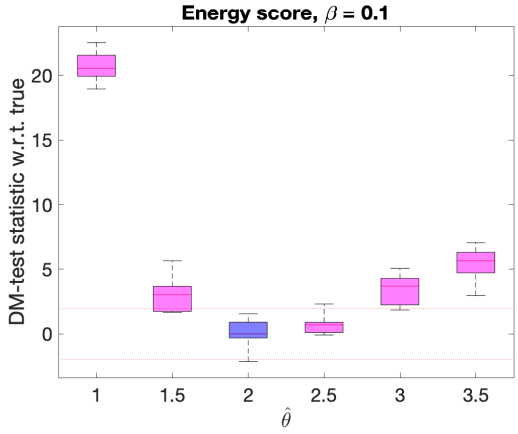
Considering the relative change in score value, see Figure 7.21, one can observe that the variogram score of order $p = 0.5$ and the Dawid-Sebastiani score are clearly able to identify the calibrated forecast. Further, the boxes corresponding to the uncalibrated forecasts separate distinctly from the boxes corresponding to the correctly reported forecasting distribution.
For the energy score we also want to study the influence of the parameter β on the discrimination ability. For a parameter $\beta < 1$ the energy score clearly performs best. Here only the boxes corresponding to the forecast with parameters $\hat{\theta} = 2$ and $\hat{\theta} = 2.5$ overlap. It also holds that the correctly specified forecasting distribution with $\hat{\theta} = 2$ yields the smallest values for the relative change in score, i.e. the energy score is able to identify the true distribution.



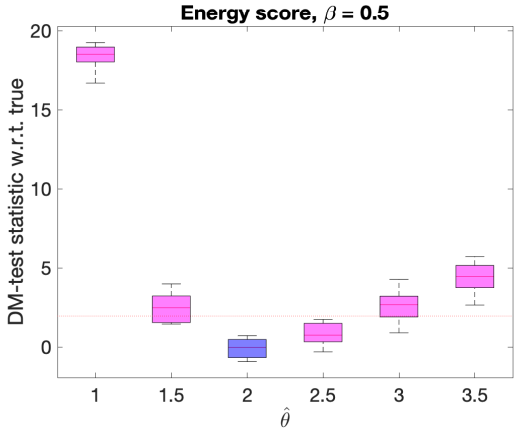
(a) Energy score with $\beta = 0.00001$



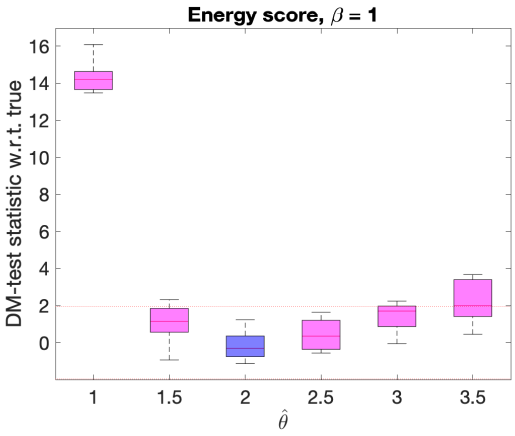
(b) Energy score with $\beta = 0.001$



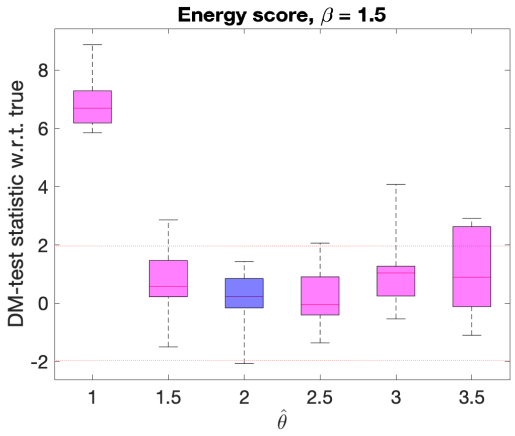
(c) Energy score with $\beta = 0.1$



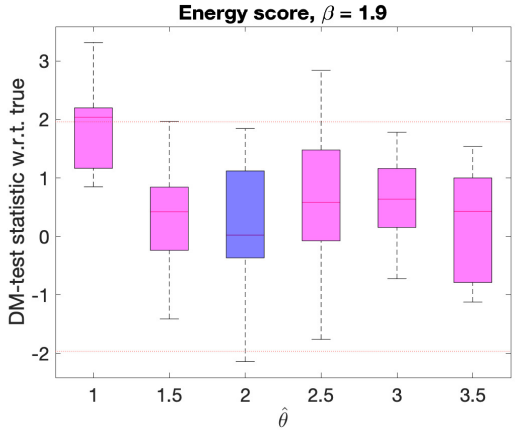
(d) Energy score with $\beta = 0.5$



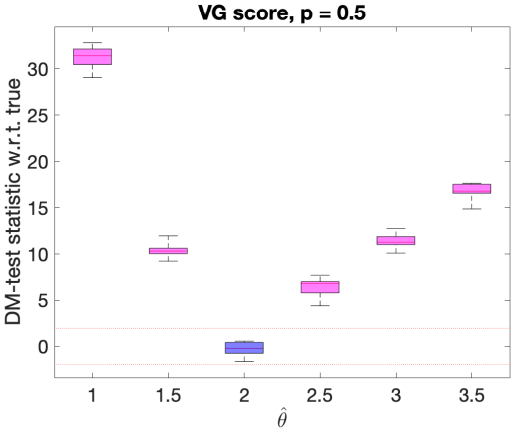
(e) Energy score with $\beta = 1$



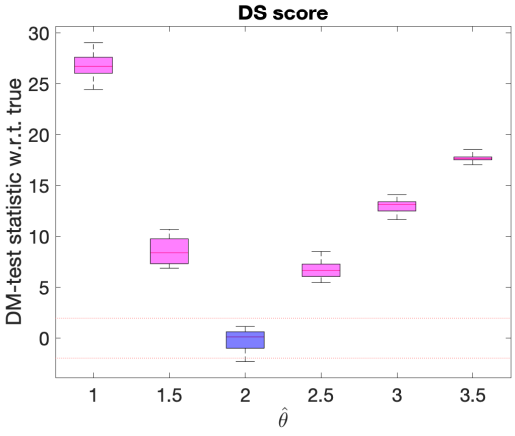
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$



(h) VG score with $p = 0.5$



(i) DS score

Figure 7.22.: As in figure 7.21, but assessed with the Diebold-Mariano test with respect to the true distribution.

However, from looking at the boxplots it does not seem that a further reduction

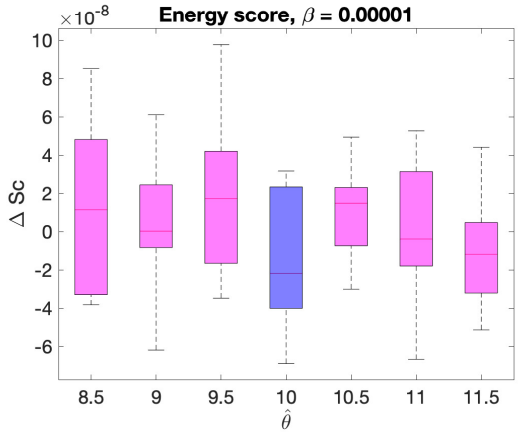
of the coefficient improves the discrimination ability of the energy score significantly. The energy score with parameters $\beta = 0.1, 0.001$, and 0.00001 yields comparable results concerning the discrimination ability of the score.
The energy score with parameters $\beta > 1$ performs noticeably worse and a greater coefficient even worsens the ability. The energy score with $\beta = 1.9$ is not able to discriminate between all forecasting distributions at all. This is not surprising as the energy score has the limiting case of the squared error in mean for $\beta = 2$ which soley depends on the first moment of the forecasting distribution.

The Diebold-Mariano test supports the preceding statements. The null hypothesis of equal predictive accuracy is rejected for all incorrectly reported forecasts when the DM-test statistic is computed with the variogram score and the Dawid-Sebastiani score. The test statistic values are all greater than 4. As a short reminder note that the null hypothesis is rejected at significance level α if the absolute value of the test statistic is greater than the $1 - \frac{\alpha}{2}$ quantile of the standard normal distributions. Again we make the standard choice of $\alpha = 5\,\%$ so the corresponding quantile is given by 1.96. Hence 1.96 and -1.96 are marked by the red dotted lines in the boxplots.
For the energy score it can be noted that for $\beta = 0.5$ some DM-test statistic values corresponding to the forecast with parameters $\hat{\theta} = 0.5$ and $\hat{\theta} = 3$ are smaller than 1.96. For $\beta = 0.1$ there are also some DM-test statistic values corresponding to the forecast with parameter $\hat{\theta} = 0.5$ smaller than 1.96, whereas for $\beta = 0.00001$ there are none.
For all parameters $\beta < 1$ the energy score is not able to discriminate between the forecasts with paramters $\hat{\theta} = 2$ and $\hat{\theta} = 2.5$.

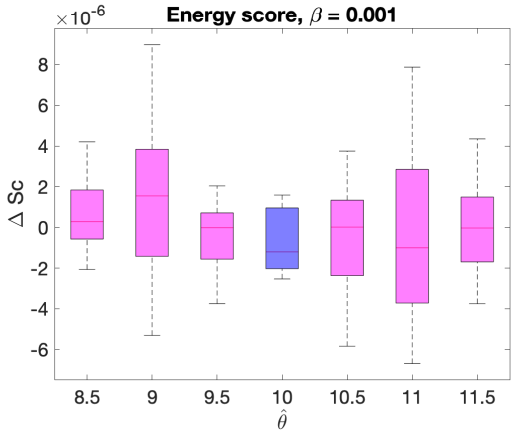
As an overall conclusion it can be stated that the variogram score clearly outperforms the energy score in this setting. Interestingly the Dawid-Sebastiani score performs similarly as the variogram score. This is surprising for the reason, that the Dawid-Sebastiani score is specifically designed for Gaussian distributions and in the setting of this simulation study the Dawid-Sebastiani score is not strictly proper.
Furthermore, this scoring rule is based on the sample mean and the sample covariance. As our variable of interest is not elliptically distributed using the covariance in such a situation might be very misleading.

In the following, we repeat the same forecasting experiment for the bivariate distribution $\mathbf{Y}$ given by the Gumbel copula with $\theta = 10$, so the upper tail dependence λ_U is much greater as in the preceding case.
Here the energy score is not able to discriminate the forecasting distributions for all parameters β. The variogram score is able to identify the calibrated forecast in the sense that the relative change in score corresponding to the calibrated forecast is the smallest. However, by looking at the DM-test statistic values the null hypothesis of equal predictive accuracy can not be rejected for parameters $\hat{\theta} = 9.5$ and $\hat{\theta} = 10.5$ in all experiments and also for $\hat{\theta} = 9$ and $\hat{\theta} = 11$ for the majority of experiments, see Figure 7.24h.

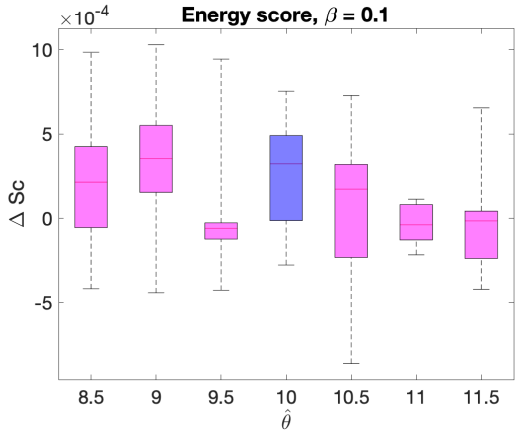
Interestingly the Dawid-Sebastiani score performs almost as good as the variogram score. The DM-test statistic values are the smallest for $\hat{\theta} = 9.5$ and $\hat{\theta} = 10$. However, also for parameters $\hat{\theta} = 9$ and $\hat{\theta} = 10.5$ the null hypothesis (falsely) cannot be rejected for all 10 forecasting experiments. Even for parameters $\hat{\theta} = 8.5$ and $\hat{\theta} = 11$ the values are partly above, partly below 1.96. Solely for $\hat{\theta} = 11.5$ the null hypothesis is always rejected, see Figure 7.24i.
Note that the overall better performance of the energy score, the variogram score, and the Dawid-Sebastiani score for the distribution given by the Gumbel copula with parameter $\theta = 2$ is obvious, as the differences in the distribution are much smaller for the greater parameter θ. For instance, the upper tail dependence of the Gumbel copula with $\theta = 2$ is $\lambda_U \approx 0.5859$ and for $\theta = 1.5$ it holds true that $\lambda_U \approx 0.412599$, whereas for $\theta = 10$ and $\theta = 9.5$ we have $\lambda_U \approx 0.9282$ and $\lambda_U \approx 0.9243$, respectively. So the differences in distribution with respect to differences in θ are much greater for a smaller parameter θ.



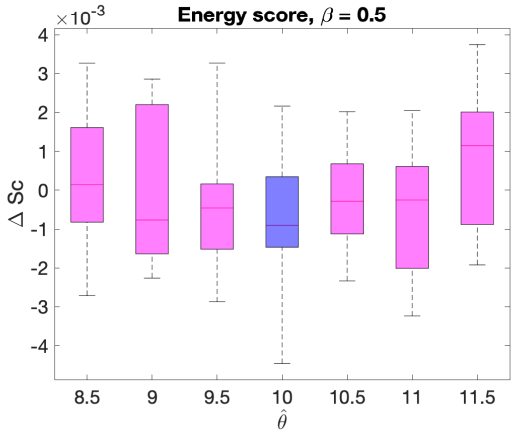
(a) Energy score with $\beta = 0.00001$



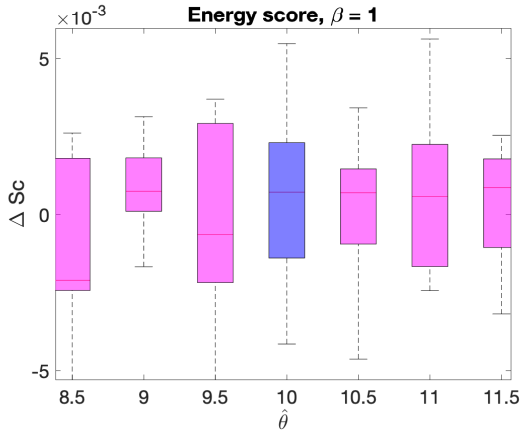
(b) Energy score with $\beta = 0.001$



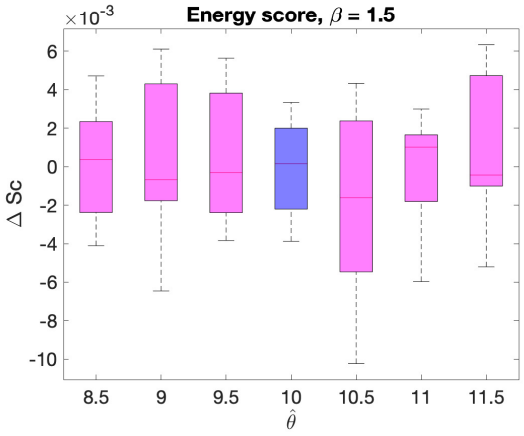
(c) Energy score with $\beta = 0.1$



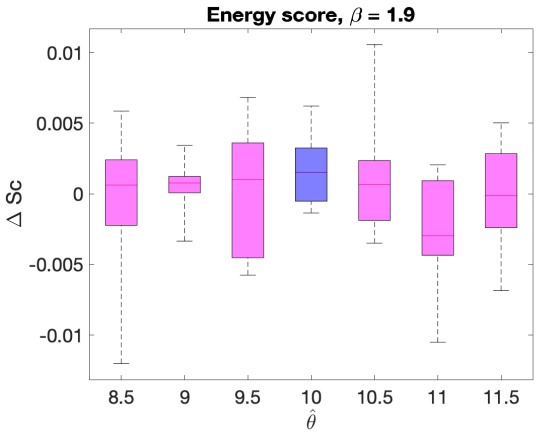
(d) Energy score with $\beta = 0.5$



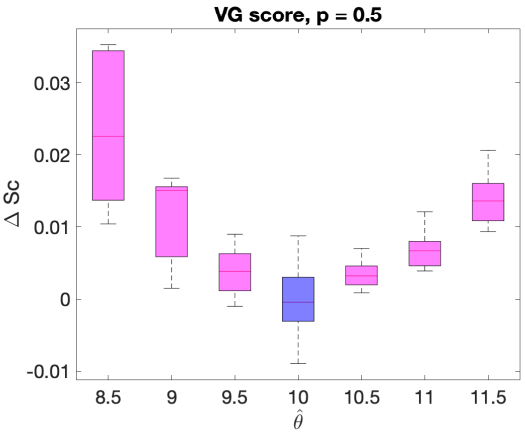
(e) Energy score with $\beta = 1$



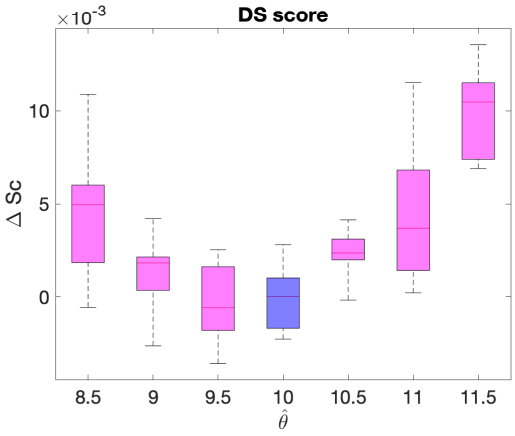
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$

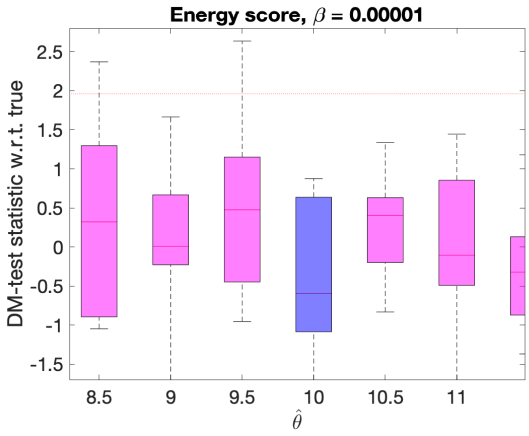


(h) VG score with $p = 0.5$

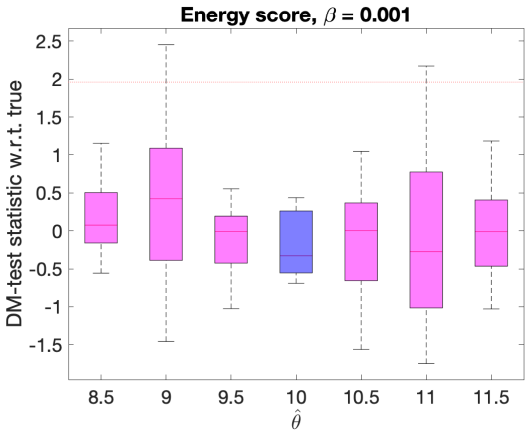


(i) DS score

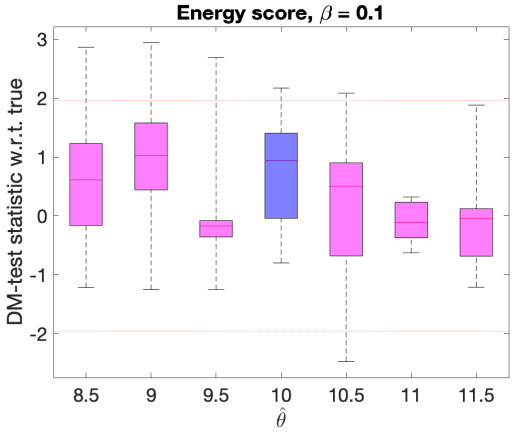
Figure 7.23.: Energy score with coefficients $\beta = 0.00001$, $\beta = 0.001$, $\beta = 0.1$, $\beta = 0.5$, $\beta = 1$, $\beta = 1.5$, and $\beta = 1.9$ as well as the variogram score with $p = 0.5$ and the Dawid-Sebastiani score assessed with the relative change in score value for an ensemble size $m = 100$. The boxplots corresponding to the forecasts with the correct and incorrect specified parameter $\hat{\theta}$ are blue and magenta, respectively. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes. The true underlying distribution is given by the bivariate Gumbel copula with parameter $\theta = 2$ and standard normal distributed marginals.



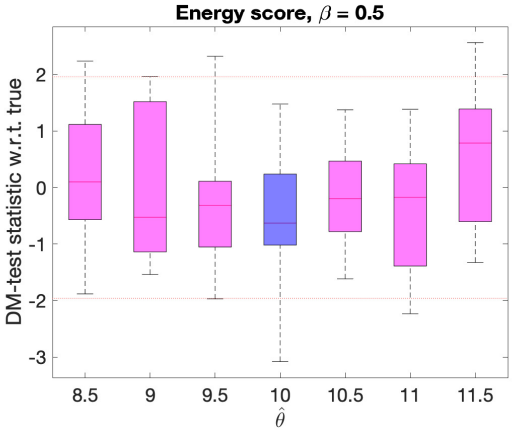
(a) Energy score with $\beta = 0.00001$



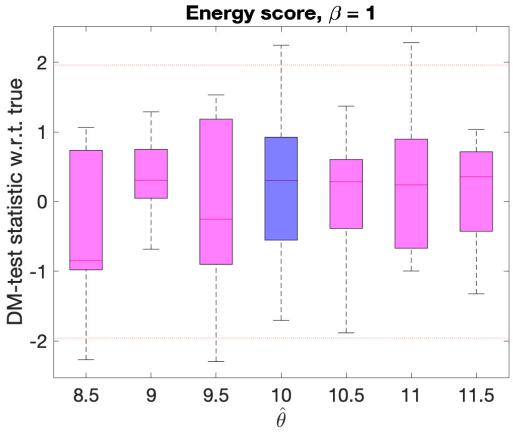
(b) Energy score with $\beta = 0.001$



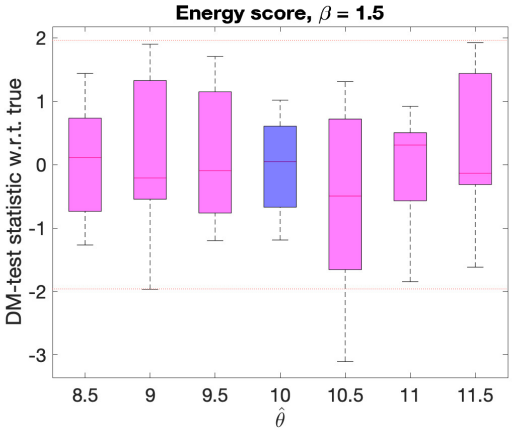
(c) Energy score with $\beta = 0.1$



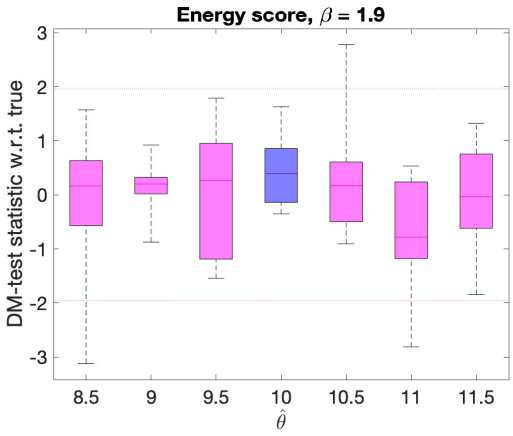
(d) Energy score with $\beta = 0.5$



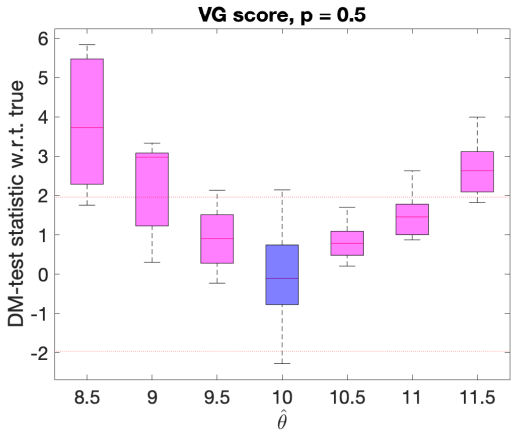
(e) Energy score with $\beta = 1$



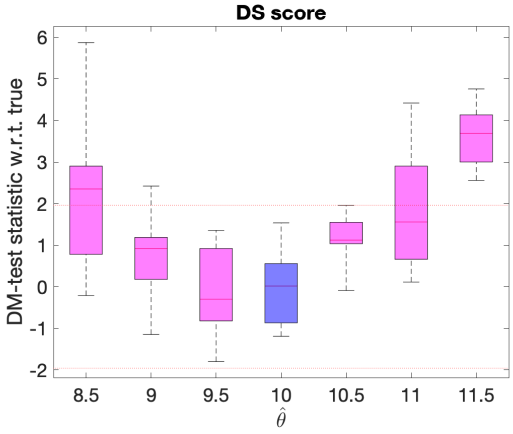
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$



(h) VG score with $p = 0.5$



(i) DS score

Figure 7.24.: As in Figure 7.23, but assessed with the Diebold-Mariano test with respect to the true distribution.

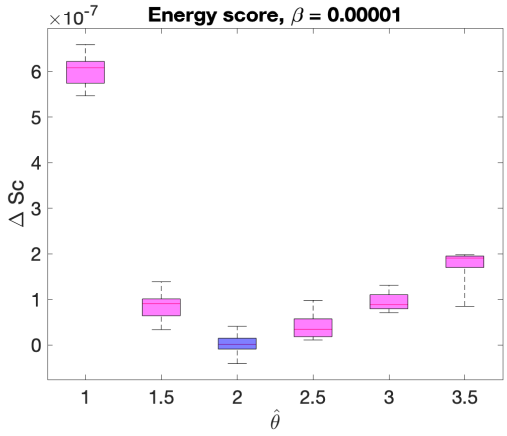
7.5.2. Beta distributed marginals

To study the influence of the marginals on the discrimination ability with respect to the dependence structure of the energy score, the variogram score and the Dawid-Sebastiani score we now assume that the marginals of the bivariate distributions $\mathbf{Y}$ are beta distributed with parameters $(\alpha, \beta) = (2, 2)$.
The density of the beta distribution is given by

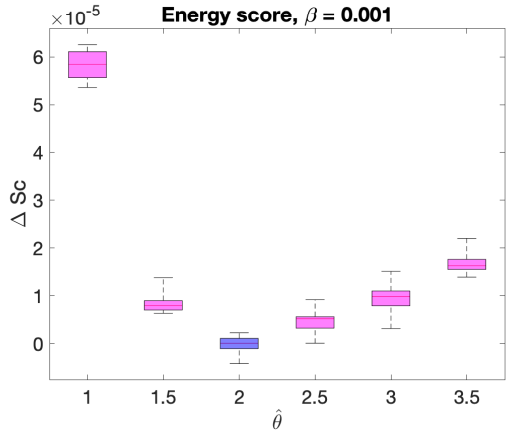
$$f(x; a, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1,$$

where $\alpha, \beta > 0$ and $\Gamma(\cdot)$ is the gamma function.
The dependency structure of the underlying distribution $\mathbf{Y}$ is given as the bivariate Gumbel copula with parameter $\theta = 2$ and the dependency structure of the forecasting distribution $\mathbf{X}$ as the Gumbel copula with parameter $\hat{\theta}$, where $\hat{\theta}$ is chosen out of the

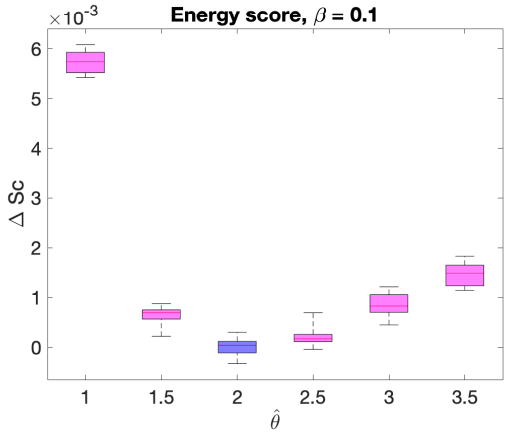
grid $\{1, 1.5, \ldots, 3.5\}$.
As above, we assess the discrimination ability of these scoring rules by the relative change in score and the Diebold-Mariano test.
Since we are interested in studying the discrimination ability of the scoring rules with respect to the dependence structure of the forecasts and underlying distribution, we assume the marginals are reported correctly.
Considering the boxplots corresponding to the relative change in score value, the outcome of this forecasting experiment with beta distributed marginals seems very similar to the preceding experiment with uniformly distributed marginals of the forecasting distribution $\mathbf{X}$.
The values of the Diebold-Mariano test statistic match this statement, so we conclude that the marginal distributions have no influence on the discrimination ability with respect to the dependence structure of the different scoring rules, if the marginals distributions are reported correctly.



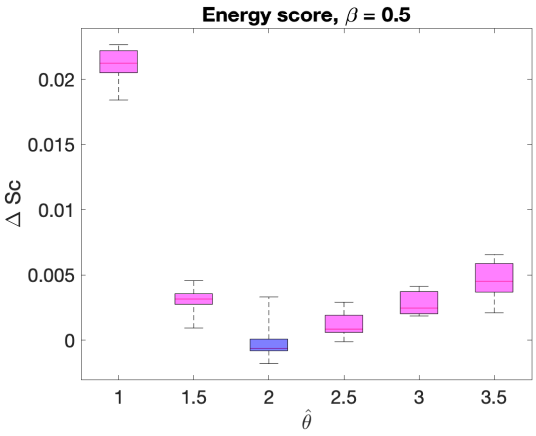
(a) Energy score with $\beta = 0.00001$



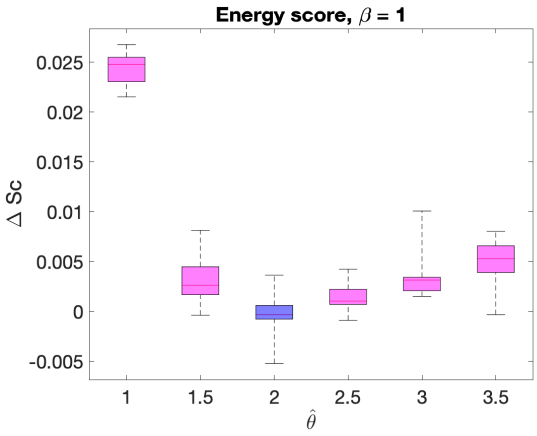
(b) Energy score with $\beta = 0.001$



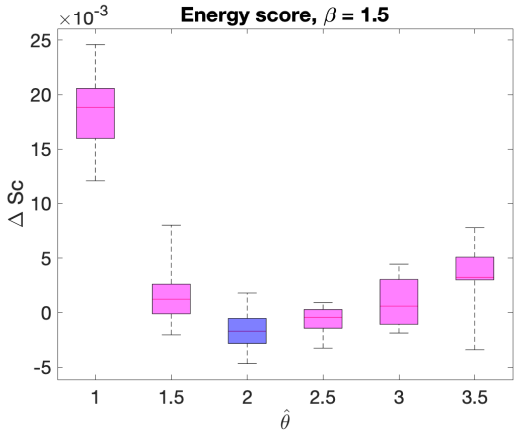
(c) Energy score with $\beta = 0.1$



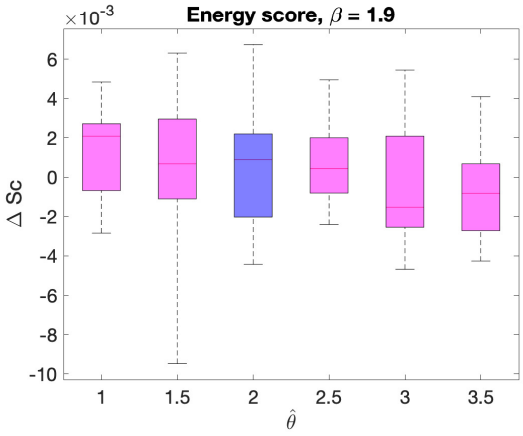
(d) Energy score with $\beta = 0.5$



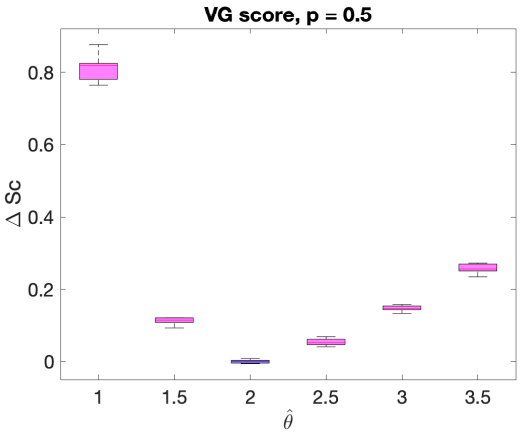
(e) Energy score with $\beta = 1$



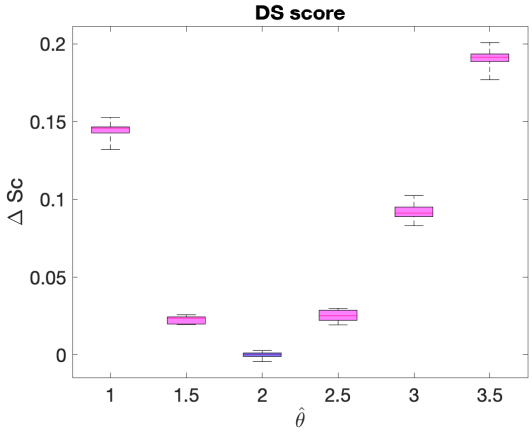
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$

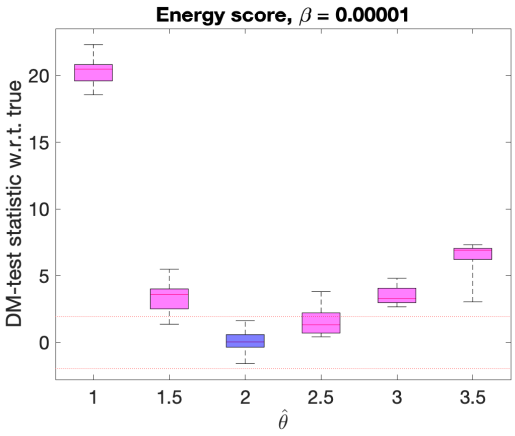


(h) VG score with $p = 0.5$

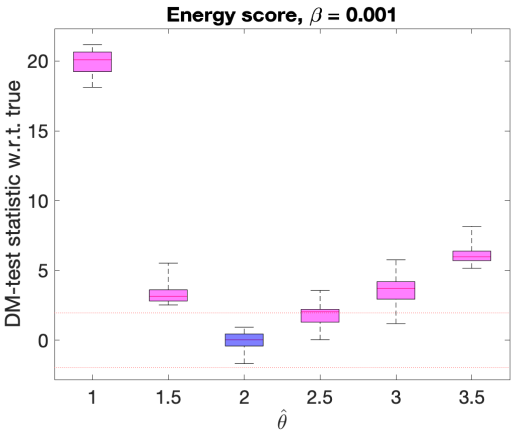


(i) DS score

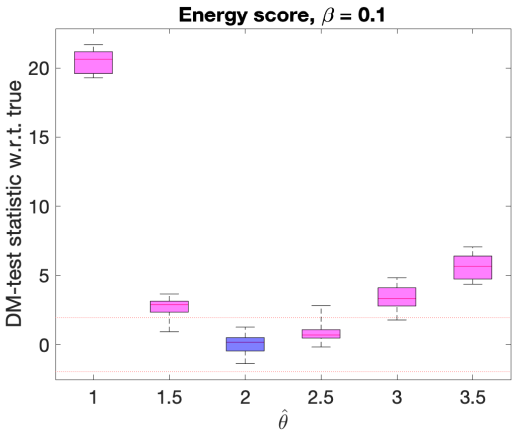
Figure 7.25.: Energy score with coefficients $\beta = 0.00001$, $\beta = 0.001$, $\beta = 0.1$, $\beta = 0.5$, $\beta = 1$, $\beta = 1.5$, and $\beta = 1.9$ as well as the variogram score with $p = 0.5$ and the Dawid-Sebastiani score assessed with the relative change in score value for an ensemble size $m = 100$. The boxplots corresponding to the forecasts with the correct and incorrect specified parameter $\hat{\theta}$ are blue and magenta, respectively. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes.



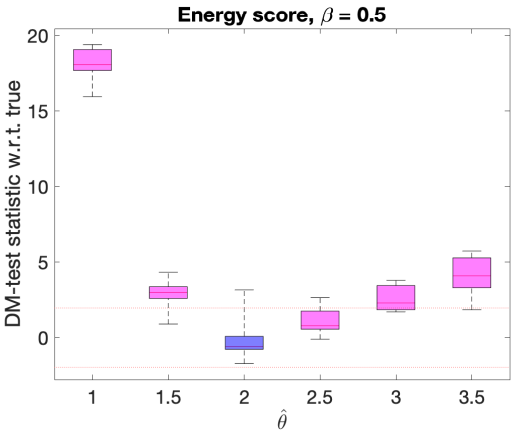
(a) Energy score with $\beta = 0.00001$



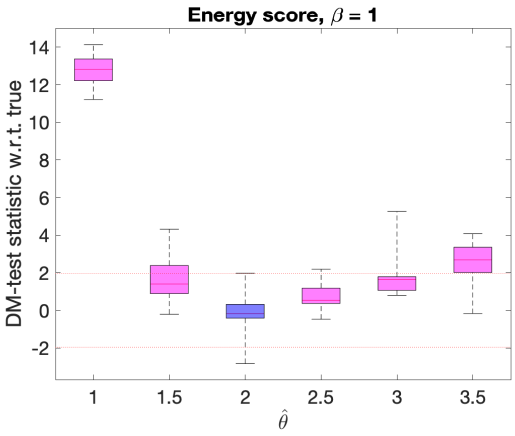
(b) Energy score with $\beta = 0.001$



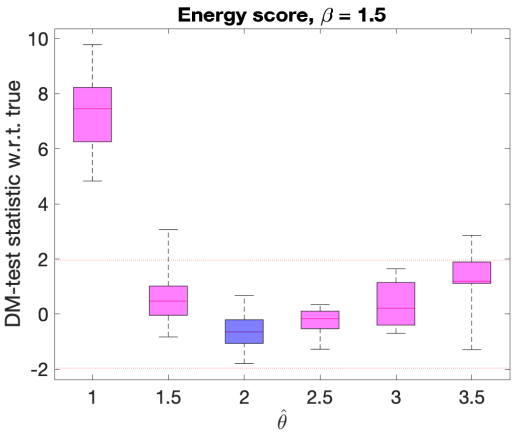
(c) Energy score with $\beta = 0.1$



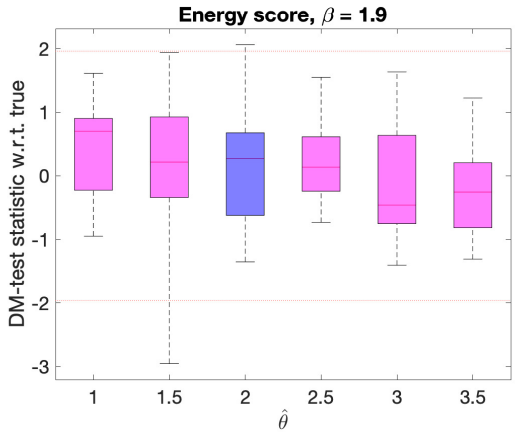
(d) Energy score with $\beta = 0.5$



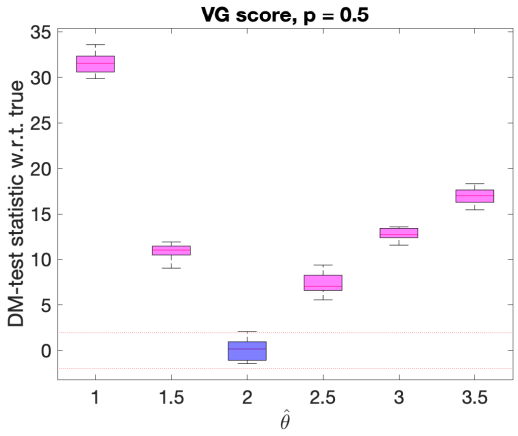
(e) Energy score with $\beta = 1$



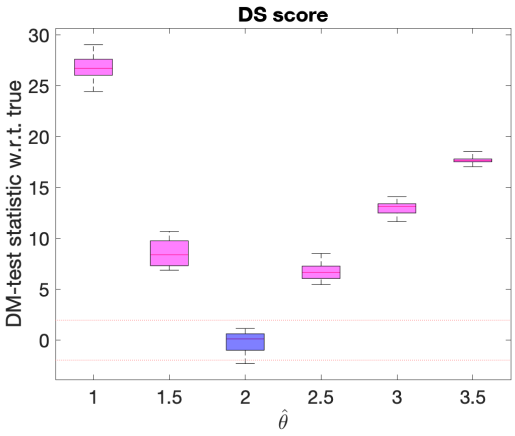
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$



(h) VG score with $p = 0.5$



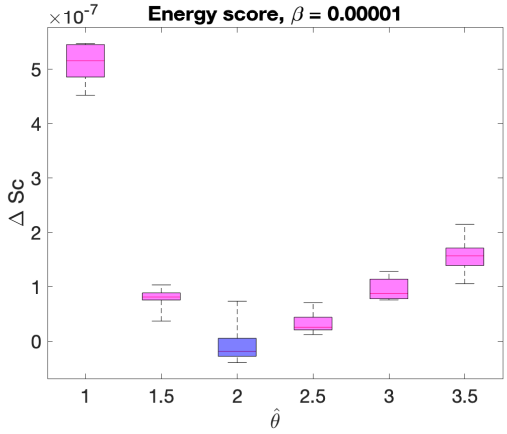
(i) DS score

Figure 7.26.: As in Figure 7.25, but assessed with the Diebold-Mariano test with respect to the true distribution.

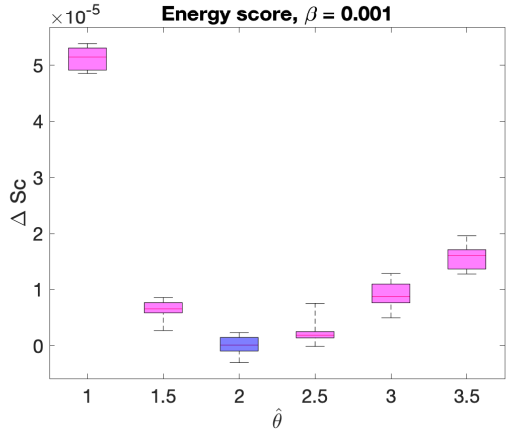
7.5.3. Normal distributed marginals

The marginal distributions in the preceding sections only take values on the interval $[0, 1]$. Therefore, we assume in this section the marginals follow a standard normal distribution and the dependence structure is given by the bivariate Gumbel copula with parameter $\theta = 2$. Furthermore, we assume that the distribution of the marginals is reported correctly and the dependeny structure of the forecast is given as the bivariate Gumbel copula with parameter $\hat{\theta}$, where $\hat{\theta}$ is chosen out of the grid $\{1, 1.5, \ldots, 3.5\}$. As above, we consider the relative change in score value and the values of the corresponding Diebold-Mariano test statistic.
Similarly to the preceding simulation studies in this section, the variogram score and the Dawid-Sebastiani score are able to identify the calibrated forecast assessed with the relative change in score as well as assessed with the Diebold-Mariano test.
Also the energy score has the same discrimination ability as in the previous simulation

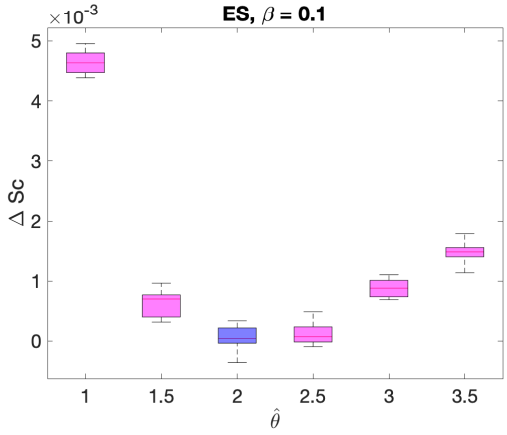
studies. Assessed with the Diebold-Mariano test the energy score with parameters $\beta = 0.00001$, $\beta = 0.001$, $\beta = 0.1$, and $\beta = 0.5$ are not able to discriminate between the calibrated forecast with $\hat{\theta} = 2$ and the uncalibrated forecast with $\hat{\theta} = 2.5$. In both cases the null hypothesis of equal predictive performance is not rejected. However, for all other forecasting distributions the null hypothesis is correctly rejected. The energy score with parameters $\beta \geq 1$ perform significantly worse.
Overall, this simulation study yields qualitatively exactly the same results as the preceeding simulation studies.
Therefore, we can conclude, that the marginal distribution does not have an influence on the discrimination ability of the considered scoring rules with respect to the interdependence structure of the forecasting distribution.



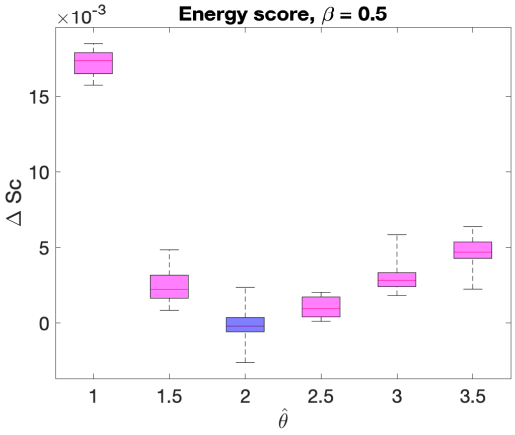
(a) Energy score with $\beta = 0.00001$



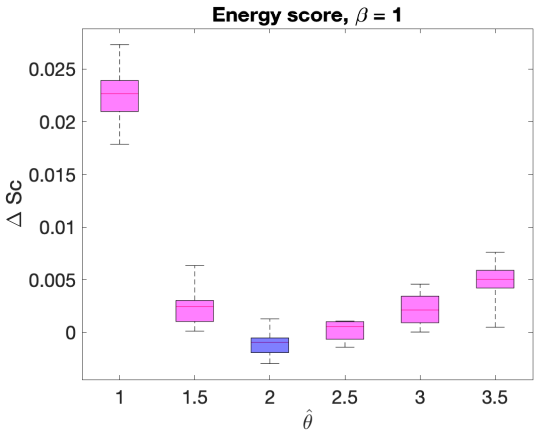
(b) Energy score with $\beta = 0.001$



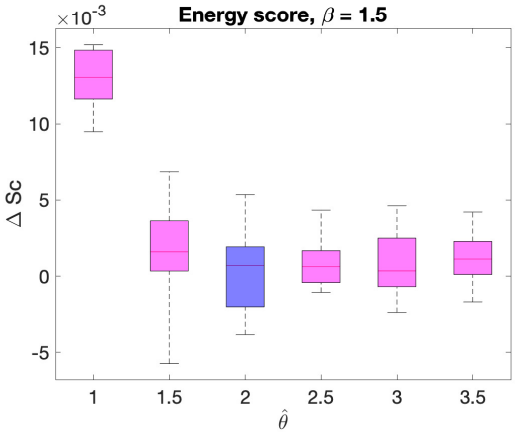
(c) Energy score with $\beta = 0.1$



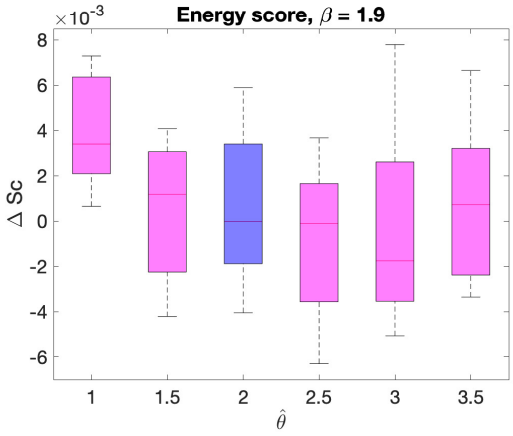
(d) Energy score with $\beta = 0.5$



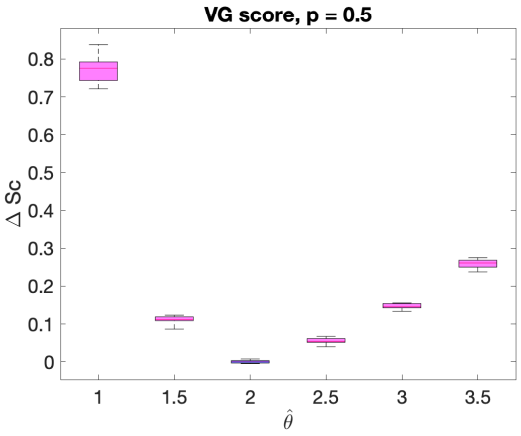
(e) Energy score with $\beta = 1$



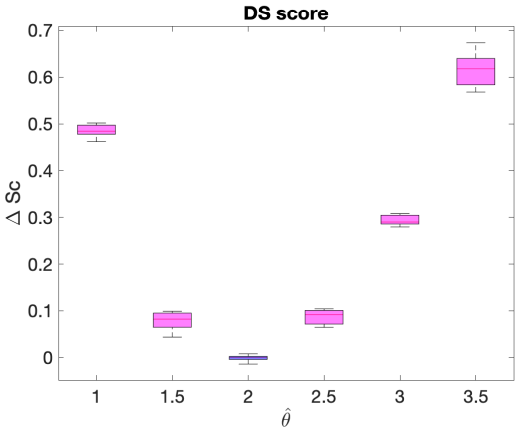
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$

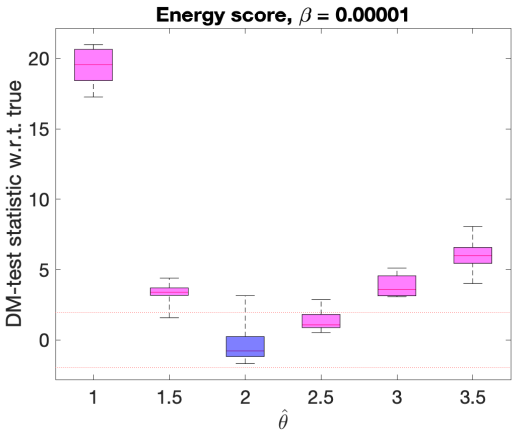


(h) VG score with $p = 0.5$

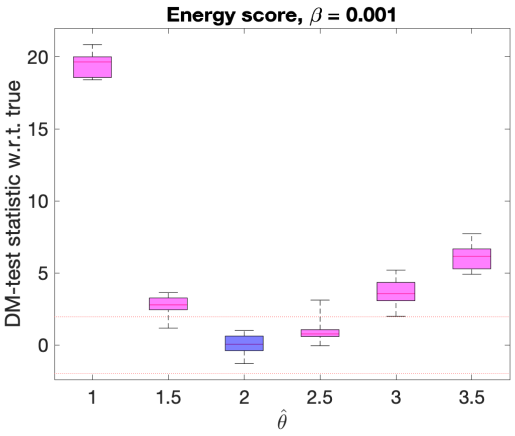


(i) DS score

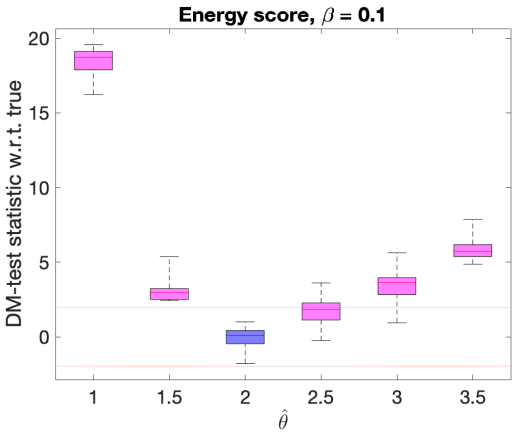
Figure 7.27.: Energy score with coefficients $\beta = 0.00001$, $\beta = 0.001$, $\beta = 0.1$, $\beta = 0.5$, $\beta = 1$, $\beta = 1.5$, and $\beta = 1.9$ as well as the variogram score with $p = 0.5$ and the Dawid-Sebastiani score assessed with the relative change in score value for an ensemble size $m = 100$. The boxplots corresponding to the forecasts with the correct and incorrect specified parameter $\hat{\theta}$ are blue and magenta, respectively. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes.



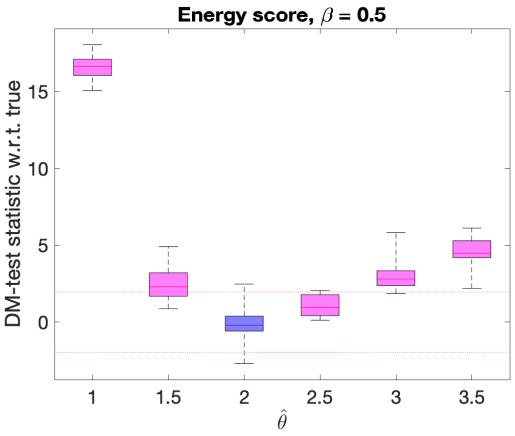
(a) Energy score with $\beta = 0.00001$



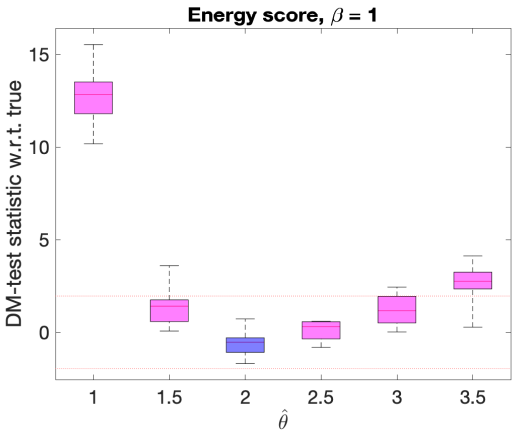
(b) Energy score with $\beta = 0.001$



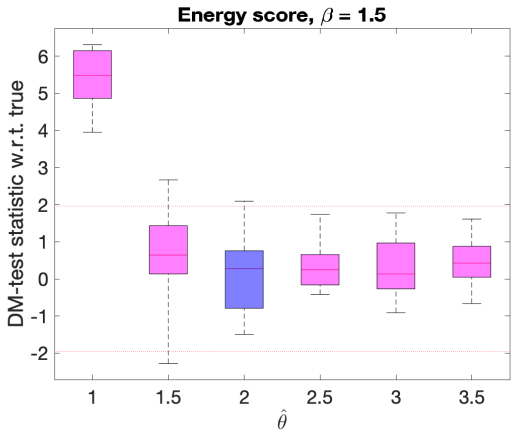
(c) Energy score with $\beta = 0.1$



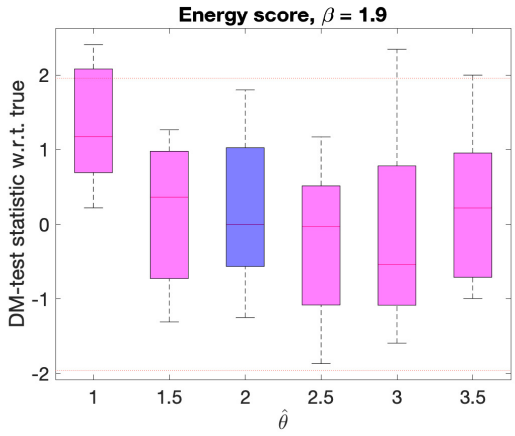
(d) Energy score with $\beta = 0.5$



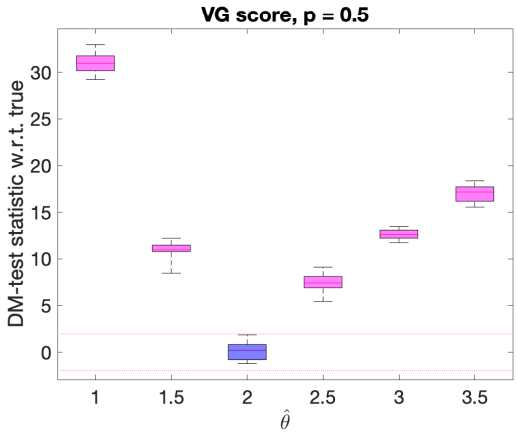
(e) Energy score with $\beta = 1$



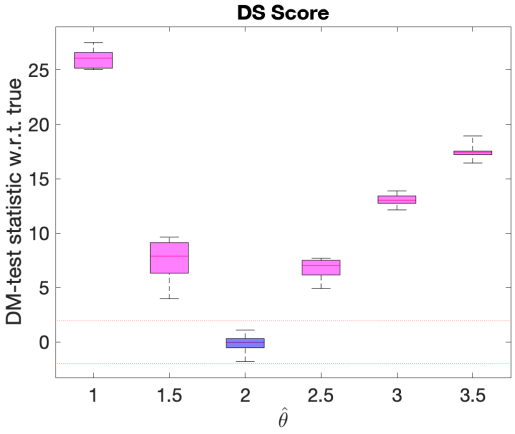
(f) Energy score with $\beta = 1.5$



(g) Energy score with $\beta = 1.9$



(h) VG score with $p = 0.5$

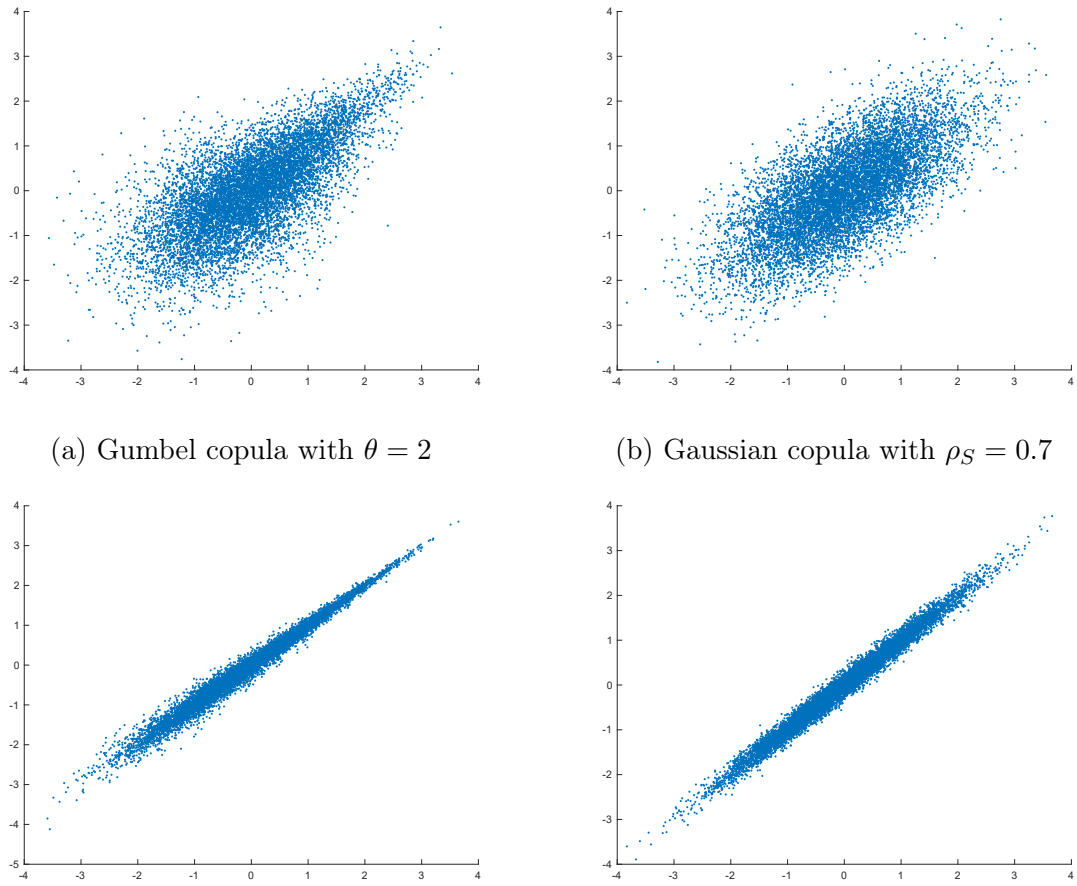


(i) DS score

Figure 7.28.: As in Figure 7.25, but assessed with the Diebold-Mariano test with respect to the true distribution.

7.5.4. Misspecified dependence model

In this section let the true underlying distribution be given by a bivariate Gumbel copula. In the following, we compare two forecasting distributions, namely the calibrated forecast and a forecast given by the bivariate Gaussian copula, so that the empirical linear correlation (Pearson correlation) of both distributions coincide.
That is, we want to find out which scoring rule assessed with the relative change in score and the Diebold-Mariano test is able to determine the correct forecast.
Note that the Pearson correlation does not satisfy all axioms for measures of multivariate concordance developed by Scarsini, see Joe (2014).



(a) Gumbel copula with $\theta = 2$

(b) Gaussian copula with $\rho_S = 0.7$

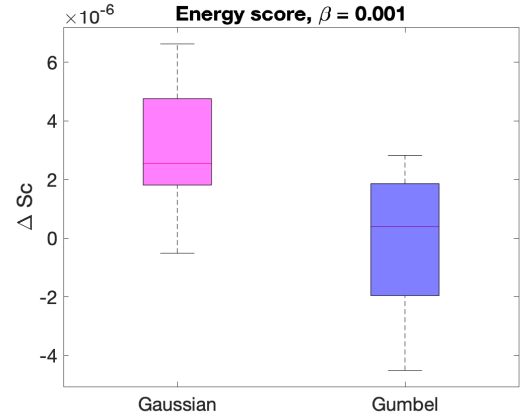
(c) Gumbel copula with $\theta = 12$

(d) Gaussian copula with $\rho_S = 0.99$

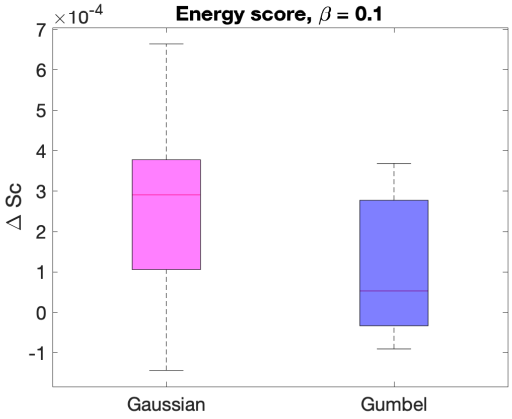
Figure 7.29.: 10.000 realizations of a Gumbel copula and a Gaussian copula with $\mathcal{N}(0,1)$ - marginals, so that the empirical linear correlation $\hat{\rho_S}$ of the Gumbel copula on the left side and the corresponding Gaussian copula on the right side match.

The distributions of the Gumbel copula and the Gaussian copula differ significantly, see Figure 7.29, even though their empirical linear correlations $\hat{\rho_S}$ approximately coincide. One can clearly observe a sharper top right corner for the Gumbel copula compared to the elliptical shape of the Gaussian copula which is known to be due to the upper tail dependence. It should be clear that the tail dependence of the Gumbel copula is greater for a greater dependence parameter θ.
Therefore, this forecasting experiment is performed for dependence parameters $\theta = 2$ and 12 of the Gumbel copula.

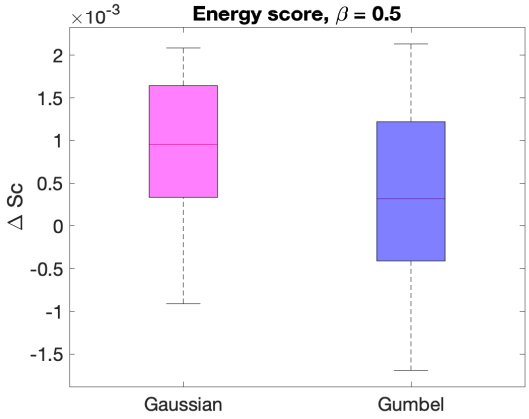
In this simulation study, the variogram score and the Dawid-Sebastiani score are able to identify the calibrated forecast. Let us first consider the variogram score. Regarding the relative change in score the boxes corresponding to the correct and incorrect specified forecast clearly separate and the median of the relative change in score is approximately 0 for the perfect forecast and approximately 0.03 for the misspecified forecast, see Figure 7.30e. The corresponding DM-test statistic values, see Figure 7.31e, match this result and the null hypothesis of equal predictive accuracy is clearly rejected when the forecast is wrongly specified as the Gaussian copula.



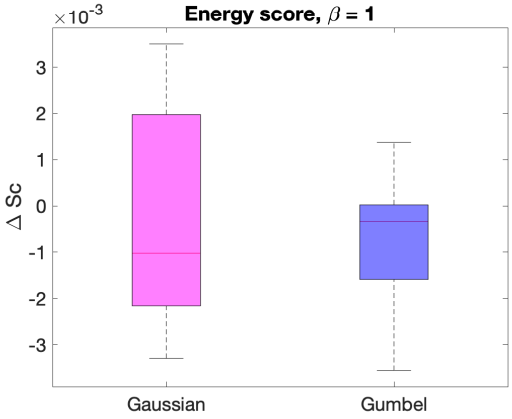
(a) Energy score with $\beta = 0.001$



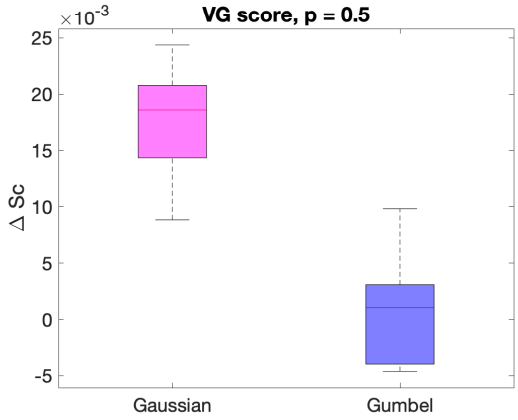
(b) Energy score with $\beta = 0.1$



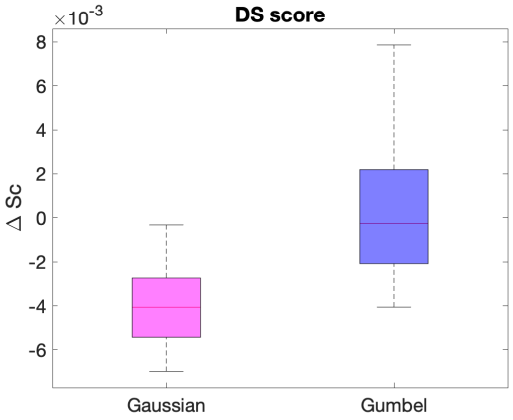
(c) Energy score with $\beta = 0.5$



(d) Energy score with $\beta = 1$



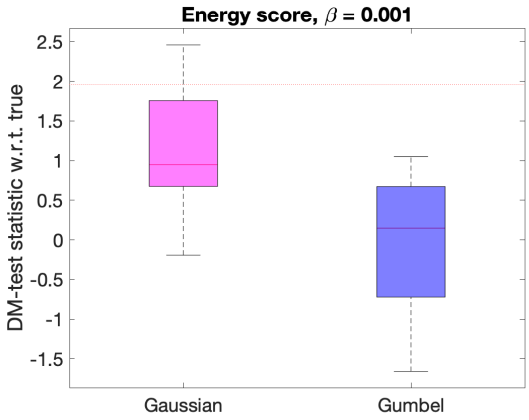
(e) VG score, $p = 0.5$



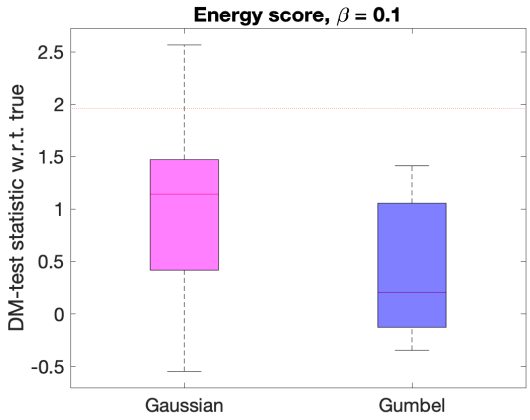
(f) DS score

Figure 7.30.: Energy score with coefficients $\beta = 0.001$, $\beta = 0.1$, $\beta = 0.5$, and $\beta = 1$ as well as the variogram score with $p = 0.5$ and the Dawid-Sebastiani score assessed with the relative change in score value for an ensemble size $m = 100$. The boxplots corresponding to the miscalibrated forecast given by a Gaussian copula and the calibrated forecast given by the Gumbel copula with dependence parameter $\theta = 2$ magenta and blue, respectively. The empirical linear correlation $\hat{\rho_S}$ of both distributions approximately match and $\hat{\rho_S} \approx 0.7$. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes.

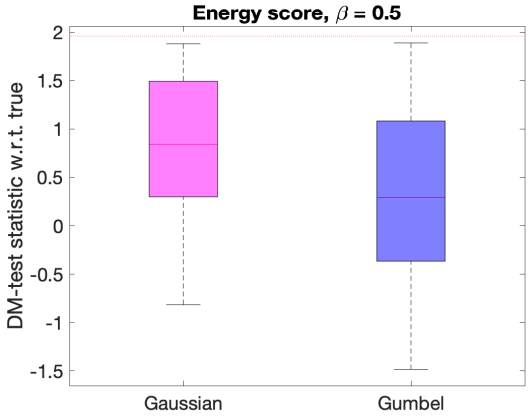
The Dawid-Sebastiani score is also able to identify the calibrated forecast, at least to a certain extent. Considering the relative change in score, the boxes corresponding to the correct and incorrect specified forecast also separate, but not als distinctly as for the variogram score. The median of the relative change in score corresponding to the perfect forecast is approximately 0, and approximately -3×10^{-3} for the misspecified forecast, see Figure 7.30f. The Dawid-Sebastiani score can take negative and positive values. Hence, the relative change in score can also take negative values. Therefore, the relative change in score has to be considered in terms of absolute values here, so the Dawid-Sebastiani score correctly favors the calibrated forecast.
The Diebold-Mariano test confirms these results. The DM-test statistic values of the perfect forecast lie all in the range from -1.96 to 1.96, so the null hypothesis is correctly not rejected. The median of DM-test statistic values corresponding to the uncalibrated forecast is approximately 2.In 5 of the 10 simulation runs the null hypothesis is correctly rejected and in the other 5 runs the null hypothesis is incorrectly not rejected, see Figure 7.31f. Only the energy score is not able to identify the calibrated forecast for all parameters β.
Further, all of these statements are true for the forecasting experiment, where the Gumbel copula is given with dependence parameter $\theta = 12$ and the corresponding linear correlation $\hat{\rho_S} \approx 0.99$, see Figure 7.32, and Figure 7.33.



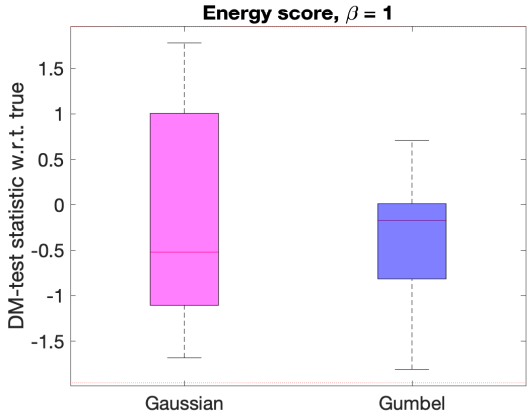
(a) Energy score with $\beta = 0.001$



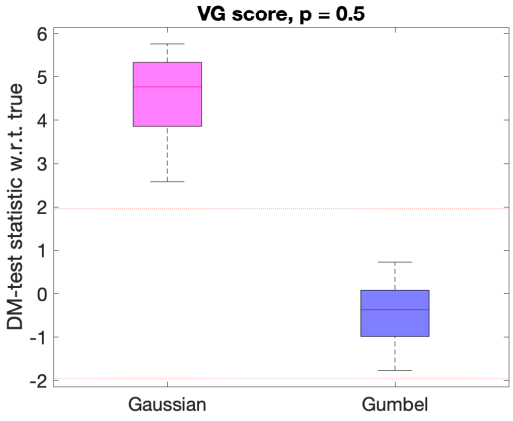
(b) Energy score with $\beta = 0.1$



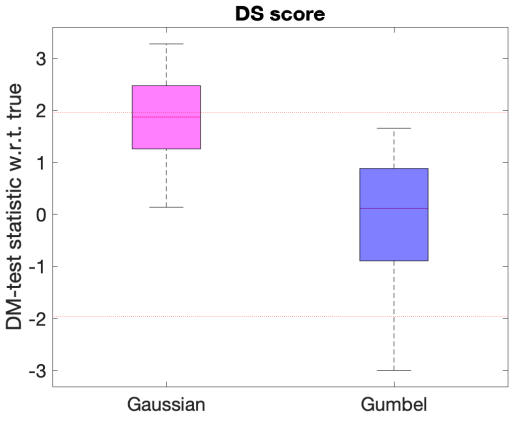
(c) Energy score with $\beta = 0.5$



(d) Energy score with $\beta = 1$

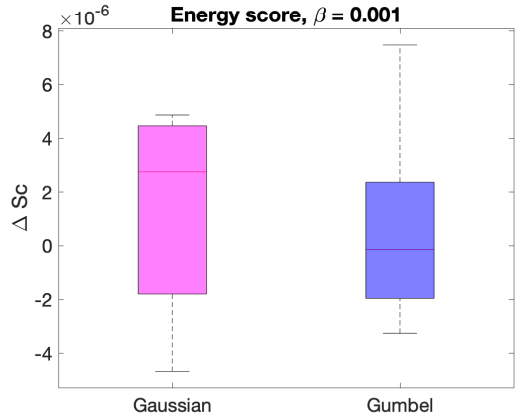


(e) VG score, $p = 0.5$

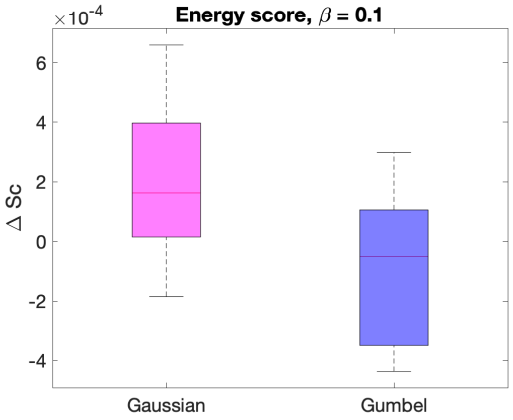


(f) DS score

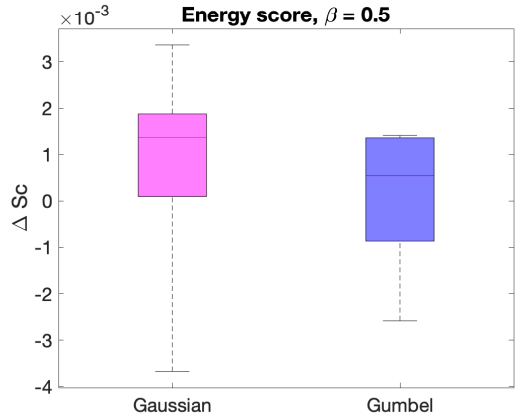
Figure 7.31.: Same as Figure 7.30, but assessed with the Diebold-Mariano test.



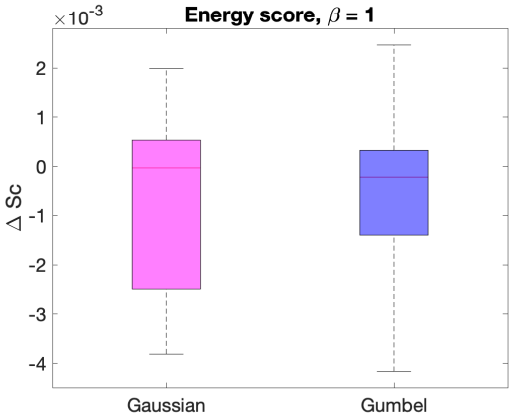
(a) Energy score with $\beta = 0.001$



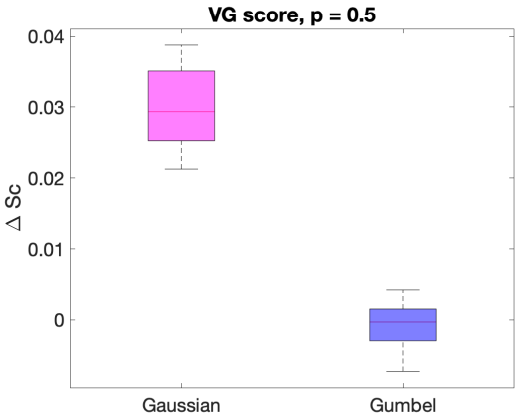
(b) Energy score with $\beta = 0.1$



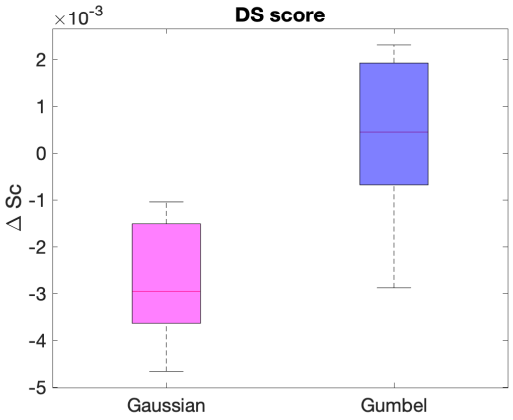
(c) Energy score with $\beta = 0.5$



(d) Energy score with $\beta = 1$

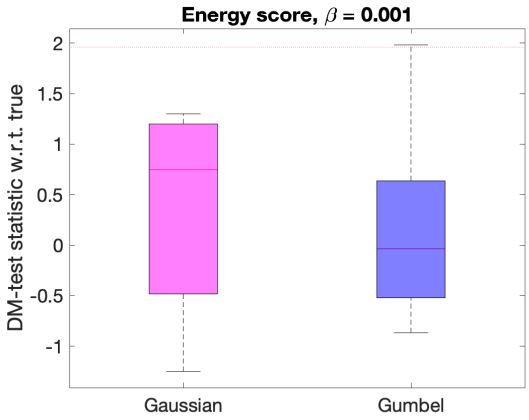


(e) VG score, $p = 0.5$

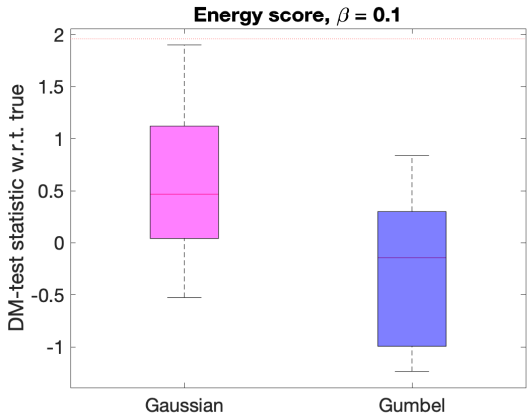


(f) DS score

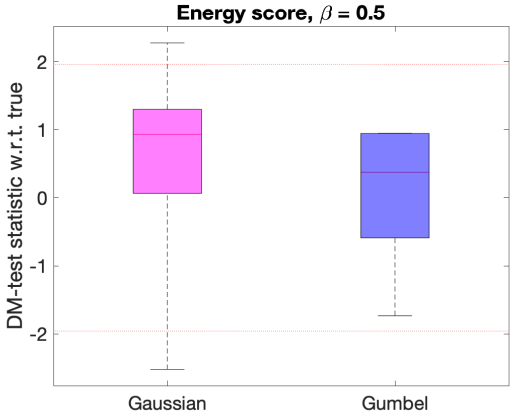
Figure 7.32.: Energy score with coefficients $\beta = 0.001$, $\beta = 0.1$, $\beta = 0.5$ and $\beta = 1$ as well as the variogram score with $p = 0.5$ and the Dawid-Sebastiani score assessed with the relative change in score value for an ensemble size $m = 100$. The boxplots corresponding to the miscalibrated forecast given by a Gaussian copula and the calibrated forecast given by the Gumbel copula with dependence parameter $\theta = 12$ magenta and blue, respectively. The empirical linear correlation $\hat{\rho}_S$ of both distributions approximately match and $\hat{\rho}_S \approx 0.99$. The boxes cover the first to third quartile of the 10 outcomes, the line shows the median, and the whiskers extend to the data extremes.



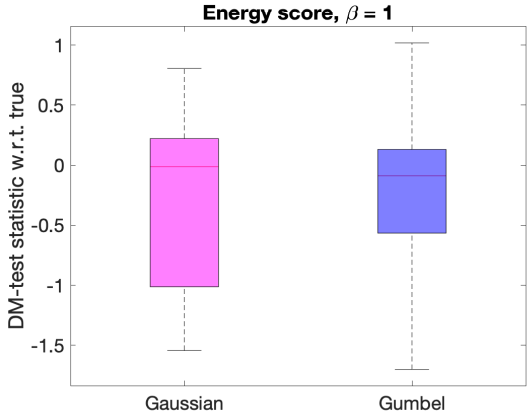
(a) Energy score with $\beta = 0.001$



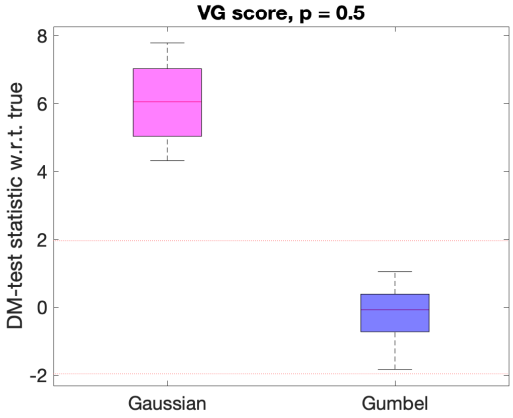
(b) Energy score with $\beta = 0.1$



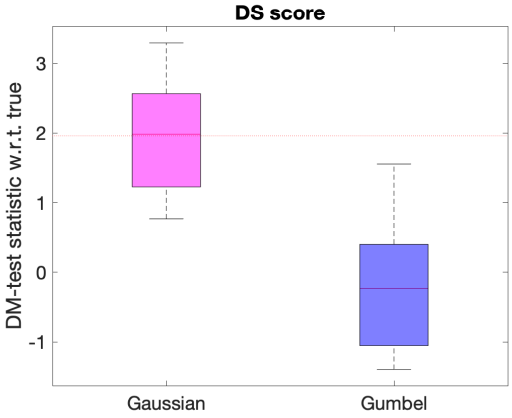
(c) Energy score with $\beta = 0.5$



(d) Energy score with $\beta = 1$



(e) VG score, $p = 0.5$



(f) DS score

Figure 7.33.: Same as Figure 7.32, but assessed with the Diebold-Mariano test.

7.5.5. Discussion of the study results

Firstly, and probably most importantly, it can be stated that the discrimination ability of all scoring rules with respect to differences in the interdependence structure does not depend on the marginal distributions. This confirms the conclusion of Ziel and Berk (2019), that the copula-scores do not have a better discrimination ability than the known original multivariate scoring rules.
Furthermore, this simulation study again confirms the good discrimination ability of the energy score when the parameter β is small, particularly, if it is approximately zero. Contrary, as β tends to 2, the energy score becomes more insensitive as it converges to the squared error in mean, which soley depends on the first moment of the forecasting distribution.

8. More on the energy distance

In almost all simulation studies the energy score is the evaluation measure which can clearly separate the true model from all alternatives. This contradicts the conclusion of Pinson and Tastu (2013) that the energy score is not suitable for evaluating differences in the interdependency structure, especially correlations.
Furthermore, the energy score is a strictly proper scoring rule, whereas the variogram score is only proper and the Dawid-Sebastiani score is strictly proper only for Gaussian distributions. Moreover, the energy score is readily applicable to ensemble forecasts, which is convenient for application. In contrast, to assess an ensemble forecast with the Dawid-Sebastiani score we have to estimate the mean and covariance matrix of the forecasting distribution first. As already stated in Section 7.4, the estimation of these parameters is rather inaccurate for small ensemble sizes and in these cases the Dawid-Sebastiani score may lead to false conclusion. As in some fields small ensemble sizes are common, this is another major drawback of this scoring rule.
So altogether the performed simulation studies allow the conclusion that the energy score should be the preferred evaluation measure for multivariate prediction. It also can be concluded, that a parameter $\beta < 1$ should be utilized, especially for evaluating differences in the interdependency structure. Note that in literature the standard choice is $\beta = 1$, even though a smaller parameter β (particularly $\beta < 1$) seems to improve the discrimination ability of the energy score.
Therefore, the energy score seems the most promising evaluation criterion so far and it is worth to study the corresponding energy distance in more detail.
Note that the energy distance allows for many statistical applications. It allows to construct a general (non-parametric) test for equality of two multivariate distributions. For instance, the resulting test for multivariate normality has a better empirical power than standard alternatives, see Szekely and Rizzo (2005). Another important application of the energy distance is the construction of distance correlations which is a general concept to characterize multivariate dependence, but not just linear dependence as measured by Pearson correlation. As a result, it allows to construct the energy test of independence which tests for multivariate independence.
The energy distance is defined as follows.

Definition 8.0.1 (Székely (2003))**.** The *generalized energy distance* between the d-dimensional independent random variables $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^{\beta} - \mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^{\beta} - \mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^{\beta} \tag{8.1}$$

where $\mathbb{E}\|\mathbf{X}\|^{\beta} < \infty$, $\mathbb{E}\|\mathbf{Y}\|^{\beta} < \infty$ for some $\beta \in (0, 2)$, $\tilde{\mathbf{X}}$ is an i.i.d copy of $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ is an i.i.d. copy of $\mathbf{Y}$.

One of the most important properties of the energy distance is, that it can be represented alternatively as a weighted L_2-distance of the characteristic functions $\hat{f}(\mathbf{t}) := \mathbb{E}(\exp(i\mathbf{t}'\mathbf{X})$ and $\hat{g}(\mathbf{t}) := \mathbb{E}(\exp(i\mathbf{t}'\mathbf{Y}))$. This means, in (8.3) we are measuring the distance in the complex plane which is isomorphic to the 2-dimensional space $\mathbb{R}^2$. Therefore, if the dimension size d is greater than 2, we measure the discrepancy in a lower dimensional space. This gives an intuition, why the energy score works so well. The following statements and proofs are essentially taken from Székely (2003).

Proposition 8.0.2. Let $\mathbf{X}$ and $\mathbf{Y}$ be independent d-dimensional random variables with characteristic functions $\hat{f}$ and $\hat{g}$. If $\mathbb{E}\|\mathbf{X}\|^\beta < \infty$ and $\mathbb{E}\|\mathbf{Y}\|^\beta < \infty$ for some $0 < \beta \leq 2$, then:

(i) For $0 < \beta < 2$,

$$\mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta - \mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - \mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta \tag{8.2}$$

$$= \frac{1}{C(d,\beta)} \int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\beta}} \mathrm{d}\mathbf{t}, \tag{8.3}$$

where

$$C(d,\beta) = 2\pi^{d/2} \frac{\Gamma(1-\beta/2)}{\beta 2^\beta \Gamma\left(\frac{d+\beta}{2}\right)}; \tag{8.4}$$

(ii)

$$\mathcal{E}^{(2)}(\mathbf{X}, \mathbf{Y}) = 2\|\mathbb{E}(\mathbf{X}) - \mathbb{E}(\mathbf{Y})\|^2.$$

Proof. Statement (ii) clearly holds. For (i), let $\overline{f(\cdot)}$ denote the complex conjugate of $f(\cdot)$. Note that

$$\begin{aligned}
|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2 &= \left(\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})\right) \overline{\left(\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})\right)} \\
&= \left(1 - \hat{f}(\mathbf{t})\overline{\hat{g}(\mathbf{t})}\right) + \left(1 - \overline{\hat{f}(\mathbf{t})}\hat{g}(\mathbf{t})\right) - \left(1 - \hat{f}(\mathbf{t})\overline{\hat{f}(\mathbf{t})}\right) - \left(1 - \hat{g}(\mathbf{t})\overline{\hat{g}(\mathbf{t})}\right) \\
&= \mathbb{E}\Big[(2 + \exp(i\mathbf{t}'(\mathbf{X} - \mathbf{Y})) - \exp(i\mathbf{t}'(\mathbf{Y} - \mathbf{X}))) \\
&\quad - \left(1 - \exp(i\mathbf{t}'(\mathbf{X} - \tilde{\mathbf{X}}))\right) - \left(1 - \exp(i\mathbf{t}'(\mathbf{Y} - \tilde{\mathbf{Y}}))\right) \Big] \\
&= \mathbb{E}\left[2\left(1 - \cos(\mathbf{t}'(\mathbf{X} - \mathbf{Y})\right) - \left(1 - \cos(\mathbf{t}'(\mathbf{X} - \tilde{\mathbf{X}})\right) - \left(1 - \cos(\mathbf{t}'(\mathbf{Y} - \tilde{\mathbf{Y}})\right)\right],
\end{aligned}$$

therefore,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\beta}} \mathrm{d}\mathbf{t} \\
&= \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{2\left(1 - \cos(\mathbf{t}'(\mathbf{X} - \mathbf{Y})\right) - \left(1 - \cos(\mathbf{t}'(\mathbf{X} - \tilde{\mathbf{X}})\right) - \left(1 - \cos(\mathbf{t}'(\mathbf{Y} - \tilde{\mathbf{Y}})\right)}{\|\mathbf{t}\|^{d+\beta}} \mathrm{d}\mathbf{t} \right].
\end{aligned}$$

Consequently, for (i) all we need to prove is the following lemma. □

Lemma 8.0.3. If $0 < \beta < 2$, it holds for all $\mathbf{x} \in \mathbb{R}^d$ that

$$\int_{\mathbb{R}^d} \frac{1 - \cos(\mathbf{t}'\mathbf{x})}{||\mathbf{t}||^{d+\beta}} \mathrm{d}\mathbf{t} = C(d,\beta)||\mathbf{x}||_d^\beta,$$

where $\mathbf{t}'\mathbf{x}$ represents the inner product, $C(d,\beta)$ is the constant (8.4) defined in the previous proposition, $\mathbf{t} \in \mathbb{R}^d$. (The intergral at $\mathbf{t} = 0$ and $\mathbf{t} = \infty$ are meant in the principal value sense: $\lim_{\epsilon \to 0} \int_{\mathbb{R}^d \setminus \{\epsilon B + \epsilon^{-1}\overline{B}\}}$, where B is the unit ball centered at 0 and $\overline{B}$ is its complement.)

Proof. Let us first consider the proof for $\beta = 1$. By applying the orthogonal transformation $\mathbf{t} \mapsto \mathbf{z} = (z_1, \ldots, z_d)$ with $z_1 = (\mathbf{t}'\mathbf{x})/||\mathbf{x}||$ and a change of variables $\mathbf{s} = ||\mathbf{x}|| \cdot \mathbf{z}$ we get

$$\int_{\mathbb{R}^d} \frac{1 - \cos(z_1||\mathbf{x}||)}{||\mathbf{z}||^{d+1}} \mathrm{d}\mathbf{z} = ||\mathbf{x}|| \int_{\mathbb{R}^d} \frac{1 - \cos(s_1)}{||\mathbf{s}||^{d+1}} \mathrm{d}\mathbf{s},$$

where $\mathbf{s} = (s_1, \ldots, s_d)$. Then

$$c_d := C(d,1) = \int_{\mathbb{R}^d} \frac{1 - \cos(s_1)}{||\mathbf{s}||^{d+1}} \mathrm{d}\mathbf{s} = \frac{\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}.$$

Notice that $2 \cdot c_d$ is the area of the unit sphere in $\mathbb{R}^{d+1}$. In the general case, where d and β both can differ from 1 more technical steps are needed. Following Prudnikov et al. (1988), we obtain by applying Formula 3.3.2.1, p. 585, Formula 2.2.4.24., p. 298, and Formula 2.5.3.13, p. 287

$$\begin{aligned} A &:= \int_{\mathbb{R}^{d-1}} \frac{\mathrm{d}z_2 \mathrm{d}z_3 \ldots \mathrm{d}z_d}{(1 + z_2^2 + z_3^2 + \cdots + z_d^2)^{\frac{d+\beta}{2}}} \\ &= \frac{\pi^{(d-1)/2}}{\Gamma\left(\frac{d-1}{2}\right)} \int_0^\infty \frac{x^{d-2}\mathrm{d}x}{(1+x^2)^{\frac{d+\beta}{2}}} = \frac{\pi^{(d-1)/2}\Gamma\left(\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{d+\beta}{2}\right)}, \end{aligned}$$

and

$$\frac{\partial}{\partial a}\left(\int_0^\infty \frac{1 - \cos(au)}{u^{1+\beta}} \mathrm{d}u\right) = a^{\beta-1} \int_0^\infty \frac{\sin v}{v^\beta} \mathrm{d}v = a^{\beta-1} \frac{\sqrt{\pi}\Gamma\left(1 - \frac{\beta}{2}\right)}{2^\beta \Gamma\left(\frac{\beta+1}{2}\right)}.$$

By introducing new variables $s_1 := z_1$ and $s_k := s_1 z_k$ for $k = 2, \ldots, d$, this yields

$$\begin{aligned} C(d,\beta) &= A \times \int_{-\infty}^\infty \frac{1 - \cos z_1}{|z_1|^{1+\beta}} \mathrm{d}z_1 \\ &= \frac{\pi^{(d-1)/2}\Gamma\left(\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{d+\beta}{2}\right)} \times \frac{2\sqrt{\pi}\Gamma\left(1 - \frac{\beta}{2}\right)}{\beta 2^\beta \Gamma\left(\frac{\beta+1}{2}\right)}, \end{aligned}$$

which was to be proved.

□

As described by Székely and Rizzo (2013), the energy distance also works for more general functions than solely $\mathbf{x} \mapsto \|\mathbf{x}\|^\beta$ in its defintion.

Proposition 8.0.4. Let ϕ be a continuous symmetric function from $\mathbb{R}^d$ to $\mathbb{R}$, and let $\mathbf{X}$ and $\mathbf{Y}$ be independent d-dimensional random variables.

(i) A necessary and sufficient condition that

$$2\mathbb{E}\phi(\mathbf{X}-\mathbf{Y}) - \mathbb{E}\phi(\mathbf{X}-\tilde{\mathbf{X}}) - \mathbb{E}\phi(\mathbf{Y}-\tilde{\mathbf{Y}}) \geq 0, \tag{8.5}$$

where $\tilde{\mathbf{X}}$ is an i.i.d. copy of $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ is an i.i.d. copy of $\mathbf{Y}$ holds for all $\mathbf{X}, \mathbf{Y}$, such that $\mathbb{E}\left[\phi(\mathbf{X}-\tilde{\mathbf{X}}) + \mathbb{E}\phi(\mathbf{Y}-\tilde{\mathbf{Y}})\right] < \infty$ is that ϕ is conditionally negative definite.

(ii) In (8.5), a necessary and sufficient condition that

$$2\mathbb{E}\phi(\mathbf{X}-\mathbf{Y}) - \mathbb{E}\phi(\mathbf{X}-\tilde{\mathbf{X}}) - \mathbb{E}\phi(\mathbf{Y}-\tilde{\mathbf{Y}}) = 0$$

if and if only $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed is that ϕ is strictly negative definite.

According to a characterization theorem of Schoenberg, a function is conditionally negative definite, continuous, symmetric, and takes the value 0 at 0 if and only if it is the negative logarithm of an infinitely divisible characteristic function, see Berg, Christensen, and Ressel (1984), Theorem 3.2.2. The functions $\phi(x) = |x|^\beta, 0 < \beta \leq 2$ correspond to infinitely divisible characteristic functions that are symmetric stable with parameter β. Note that in the case $\phi(x) = |x|^2$ we have conditional negative definiteness, but not strict conditional negative definiteness.
Other examples include $x \mapsto \log(1 + |x|^2)$, which corresponds to the characteristic function of the Laplace distribution. In Baringhaus and Franz (2010) other examples of strictly negative functions ϕ are given. So other possible choices are for example $\phi(x) = 1 - \exp(|x|^2/2)$, $\phi(x) = |x|/2$ or $\phi(x) = |x|^2/(1 + |x|^2)$.
Thus, according to Proposition 8.0.4, the energy score can be generalized as we can utilize different functions than $x \mapsto |x|^\beta$.

9. Limiting cases of the energy score

In this section we consider the limiting cases of the energy score, as we found in the previous simulation studies that the discrimination ability of the energy score improves as $\beta \to 0$ and worsens as $\beta \to 2$. As a measure for the discrimination ability we evaluate the score values of the forecasting distributions with the Diebold-Mariano test w.r.t. the true underlying distributions.
The case $\beta \to 2$ is trivial, as the energy score reduces to the squared error $\mathrm{ES}_2 = \|\mathbb{E}(\mathbf{X}) - \mathbf{y}\|^2$, which is just a proper and not a strictly proper scoring rule with respect to the class $\mathcal{P}_2$ of Borel probability measures on $\mathbb{R}^d$ such that $\mathbb{E}\,\|\mathbf{X}\|^2$ is finite.
ES_2 depends only on the mean of the forecasting distribution, which explains the lack of sensitivity of the energy score with $\beta \to 2$ particularly with respect to errors in correlation.

So let us now turn to the case $\beta \to 0$. As an illustrating example, we consider the following forecasting experiment. Let the forecasting distribution be given by a bivariate Gaussian distribution with mean $\mu = (0, 0)$ and covariance structure

$$\hat{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

The true underlying distribution is also as a bivariate Gaussian distribution with mean $\mu = (0, 0)$ and covariance structure

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The sample size is $m = 2^{10}$ and the experiment is repeated $N = 2^{10}$ times. For each experiment we calculate the corresponding energy score values for all parameters β out of the grid $\{0.1, \ldots, 1.9\}$.
The resulting score values are assessed with the Diebold-Mariano test, see Figure 9.1. Given that the predictive distribution does not match the true underlying distribution, a higher value of the Diebold-Mariano test statistic indicates a better discrimination ability of the corresponding scoring rule with which the probabilistic forecast was evaluated.

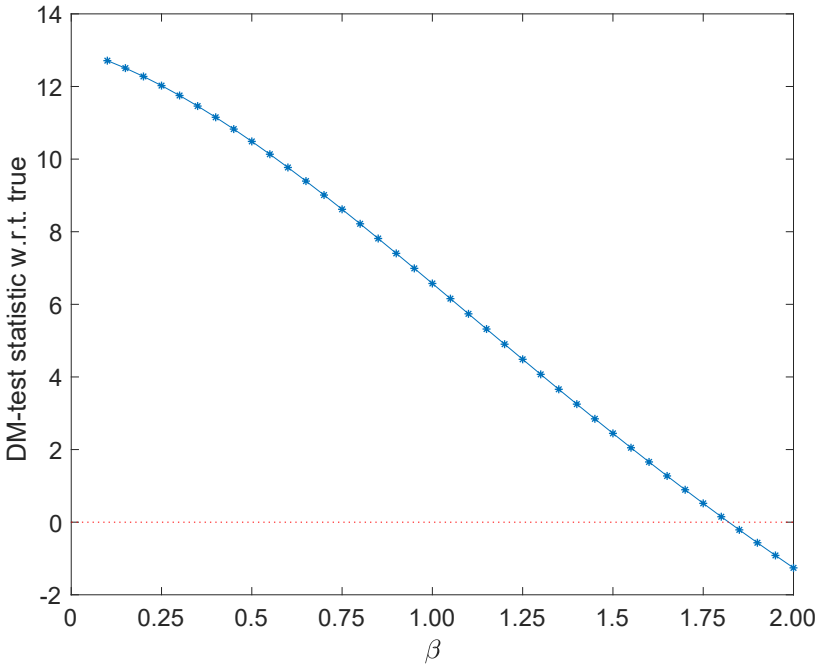


Figure 9.1.: Values of the Diebold-Mariano test statistic corresponding to the energy score values for different parameters β.

As already stated above, the discrimination ability of the energy score improves as the parameter β gets smaller. So the question arises, if there exists a strictly proper scoring rule to which ES_β converges analogous to the case where $\beta \to 2$.

To answer this, we firstly consider the corresponding divergence function of the energy score, namely the energy distance. Remember the energy distance has the following two representations:

- $\mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta - \mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - \mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta$,
  where $\mathbf{X}$ and $\mathbf{Y}$ are d-dimensional independent random variables with $\mathbb{E}\|\mathbf{X}\|^\beta < \infty$, $\mathbb{E}\|\mathbf{Y}\|^\beta < \infty$ for some $\beta \in (0, 2)$ and $\tilde{\mathbf{X}}$ is an i.i.d copy of $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ is an i.i.d. copy of $\mathbf{Y}$
- $\mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = \frac{\beta 2^\beta \Gamma\left(\frac{d+\beta}{2}\right)}{2\pi^{d/2}\Gamma\left(1-\frac{\beta}{2}\right)} \int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t}) - \hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\beta}} \mathrm{d}\mathbf{t}$.

Using the definition of the energy distance, we can set up the following lemma.

Lemma 9.0.1. Let $\mathbf{X}$ and $\mathbf{Y}$ be independent d-dimensional random variables, $\mathbb{E}\|\mathbf{X}\|^\beta < \infty$, $\mathbb{E}\|\mathbf{Y}\|^\beta < \infty$ for some $\beta \in (0, 2)$.
If $\mathbf{X}$ and $\mathbf{Y}$ are continuous random variables it holds that

$$\lim_{\beta \to 0} \mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = 0. \tag{9.1}$$

If $\mathbf{X}$ and $\mathbf{Y}$ are discrete random variables both taking values on the same set $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$, then

$$\lim_{\beta \to 0} \mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = \sum_i (p_i - q_i)^2, \tag{9.2}$$

where $p_i := \mathbb{P}(\mathbf{X} = \mathbf{x}_i)$ and $q_i := \mathbb{P}(\mathbf{Y} = \mathbf{x}_i)$:

Proof. It holds that

$$\lim_{\beta \to 0} \|\mathbf{X} - \mathbf{Y}\|^\beta = \mathbb{1}_{[\mathbf{X} \neq \mathbf{Y}]}.$$

Therefore, we have

$$\begin{aligned}
&\lim_{\beta \to 0} \mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = \lim_{\beta \to 0} 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|^\beta - \mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - \mathbb{E}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta \\
&= 2 \cdot \mathbb{E}\left(\mathbb{1}_{\{\mathbf{X} \neq \mathbf{Y}\}}\right) - \mathbb{E}\left(\mathbb{1}_{\{\mathbf{X} \neq \tilde{\mathbf{X}}\}}\right) - \mathbb{E}\left(\mathbb{1}_{\{\mathbf{Y} \neq \tilde{\mathbf{Y}}\}}\right) \\
&= 2 \cdot \left(1 - \mathbb{E}\left(\mathbb{1}_{\{\mathbf{X} = \mathbf{Y}\}}\right)\right) - \left(1 - \mathbb{E}\left(\mathbb{1}_{\{\mathbf{X} = \tilde{\mathbf{X}}\}}\right)\right) - \left(1 - \mathbb{E}\left(\mathbb{1}_{\{\mathbf{Y} = \tilde{\mathbf{Y}}\}}\right)\right) \\
&= \mathbb{E}\left(\mathbb{1}_{\{\mathbf{X} = \tilde{\mathbf{X}}\}}\right) + \mathbb{E}\left(\mathbb{1}_{\{\mathbf{Y} = \tilde{\mathbf{Y}}\}}\right) - 2 \cdot \mathbb{E}\left(\mathbb{1}_{\{\mathbf{X} = \mathbf{Y}\}}\right) \\
&= \int \mathbb{1}_{\{\mathbf{x} = \tilde{\mathbf{x}}\}} F(\mathrm{d}\mathbf{x}) F(\mathrm{d}\tilde{\mathbf{x}}) + \int \mathbb{1}_{\{\mathbf{y} = \tilde{\mathbf{y}}\}} G(\mathrm{d}\mathbf{y}) G(\mathrm{d}\tilde{\mathbf{y}}) - 2 \cdot \int \mathbb{1}_{\{\mathbf{x} = \mathbf{y}\}} F(\mathrm{d}\mathbf{x}) G(\mathrm{d}\mathbf{y}),
\end{aligned}$$

where F denotes the distribution of $\mathbf{X}$ and G the distribution of $\mathbf{Y}$.
In case of continuous distributions the random variables only match on a null set. Therefore, the limit here is 0.
Now let $\mathbf{X}$ and $\mathbf{Y}$ be discrete distributed. We assume $\mathbf{X}$ and $\mathbf{Y}$ can take the values $\mathbf{x}_1, \mathbf{x}_2, \dots$ and the corresponding probabilities are denoted as $p_i := P(\mathbf{X} = \mathbf{x}_i)$ and $q_i := P(\mathbf{Y} = \mathbf{x}_i)$.
It follows that

$$\begin{aligned}
&\int \mathbb{1}_{[\mathbf{x} = \tilde{\mathbf{x}}]} F(\mathrm{d}\mathbf{x}) F(\mathrm{d}\tilde{\mathbf{x}}) + \int \mathbb{1}_{[\mathbf{y} = \tilde{\mathbf{y}}]} G(\mathrm{d}\mathbf{y}) G(\mathrm{d}\tilde{\mathbf{y}}) - 2 \cdot \int \mathbb{1}_{[\mathbf{x} = \mathbf{y}]} F(\mathrm{d}\mathbf{x}) G(\mathrm{d}\mathbf{y}) \\
&= \sum_i p_i^2 + \sum_i q_i^2 - 2 \sum_i p_i q_i \\
&= \sum_i (p_i - q_i)^2.
\end{aligned}$$

□

Remark 9.0.2. Note that in the discrete case the limit of the energy score corresponds to the divergence function of the Brier score, see Gneiting and Raftery (2007), which is a strictly proper scoring rule with respect to the class of discrete probability distributions $\mathcal{P}_n := \{(p_1, \dots, p_n) : p_1, \dots, p_n \geq 0, p_1 + \dots + p_n = 1\}$ on a finite discrete sample space $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.

9.1. Scaled limit of the energy score

As a motivation of this section, we consider representation (8.0.2) of the energy distance. When we take the limit of this term it obviously has to match the previous results.
It holds that

$$\lim_{\beta\to 0} \frac{\beta 2^\beta \Gamma\left(\frac{d+\beta}{2}\right)}{2\pi^{d/2}\Gamma\left(1-\frac{\beta}{2}\right)} \int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t})-\hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\beta}} \mathrm{d}\mathbf{t}$$
$$= \left(\lim_{\beta\to 0} \frac{\beta 2^\beta \Gamma\left(\frac{d+\beta}{2}\right)}{2\pi^{d/2}\Gamma\left(1-\frac{\beta}{2}\right)}\right) \cdot \left(\lim_{\beta\to 0} \int_{\mathbb{R}^d} \frac{|\hat{f}(\mathbf{t})-\hat{g}(\mathbf{t})|^2}{\|\mathbf{t}\|^{d+\beta}} \mathrm{d}\mathbf{t}\right),$$

if the limits are finite.
It can easily be seen that the first term converges to 0 regardless of whether the random variables are discrete or continuously distributed.
For the second expression, however, we have to distinguish between discrete and continuous distributions. To match the previous results, it can be conjectured that this term tends to a constant in case of continuous distributions and it does not converge for discrete distributions.
Hence, the assumption follows that by scaling the energy distance by the factor $1/\beta$ it yields a limit not equal to zero as β tends to 0.
Therefore, in the following, we consider the term

$$\mathcal{E}_{sc}^{(\beta)}(\mathbf{X},\mathbf{Y}) := \frac{1}{\beta}\cdot\mathcal{E}^{(\beta)}(\mathbf{X},\mathbf{Y}).$$

Before we consider the limit of this scaled energy distance, the following technical steps are needed.

Lemma 9.1.1. The function $g_\beta(x) = x - 1 - \frac{x^\beta - 1}{\beta}$ with domain $(0,\infty)$ is non-negative for $0 < \beta < 1$.

Proof. We consider the derivate of $g_\beta(x)$

$$\frac{\partial}{\partial x} g_\beta(x) = 1 - x^{\beta-1}$$

which is zero only if $x = 1$. Further, it holds that

$$\frac{\partial^2}{\partial x^2} g_\beta(x) = (1-\beta)x^{\beta-2}.$$

As

$$\frac{\partial^2}{\partial x^2} g_\beta(1) > 0,$$

the function $g_\beta(x)$ admits a minimum at $x = 1$. As $\frac{\partial}{\partial x} g_\beta(x) > 0$ for $x > 1$ and $\frac{\partial}{\partial x} g_\beta(x) < 0$ for $0 < x < 1$ the function is monotone decreasing for $0 < x < 1$ and monotone increasing for $x > 1$. Therefore, the attained minimum in $x = 0$ is global. Therefore the function is non-negative. □

Lemma 9.1.2. For $\beta \to 0$ the family of functions g_β defined in Lemma 9.1.1 is monotone in β with limit

$$\lim_{\beta \to 0} g_\beta(x) = x - 1 - \log(x).$$

Proof. It clearly holds that

$$\lim_{\beta \to 0} \frac{x^\beta - 1}{\beta} = \log(x),$$

so the limit

$$\lim_{\beta \to 0} g_\beta(x) = x - 1 - \log(x)$$

directly follows.
Next we want to show that sequence $g_\beta(\cdot)$ is monotone increasing as $\beta \to 0$. We assume $\beta_1 < \beta_2$ and define the function

$$d(x) = g_{\beta_1}(x) - g_{\beta_2}(c) = \frac{x^{\beta_2} - 1}{\beta_2} - \frac{x^{\beta_1} - 1}{\beta_1}.$$

We have to show that the function $d(x)$ is non-negative for all $x > 0$. It holds that

$$\frac{\partial}{\partial x} d(x) = x^{\beta_2 - 1} - x^{\beta_1 - 1} = 0$$

only if $x = 1$. As it further holds that

$$\frac{\partial^2}{\partial x^2} d(x) = \frac{(\beta_2 - 1)x^{\beta_2} - (\beta_1 - 1)x^{\beta_1}}{x^2},$$

it follows that

$$\frac{\partial^2}{\partial x^2} d(1) = \beta_2 - \beta_1 > 0.$$

Therefore, the function d attains its minimum at $x = 1$. Further, it holds that $\frac{\partial}{\partial x} d(x) > 0$ for $x > 1$ and $\frac{\partial}{\partial x} d(x) < 0$ for $0 < x < 1$. Thus, the minimum at $x = 1$ is global. Furthermore, it holds that $d(1) = 0$. Therefore, $d(x)$ is non-negative.
Therefore, the sequence $g_\beta(\cdot)$ is monotone increasing as $\beta \to 0$. □

Utilizing Lemma 9.1.1 and Lemma 9.1.2, the subsequent theorem follows.

Theorem 9.1.3. Let $\mathbf{X}$ and $\mathbf{Y}$ be independent d-dimensional random variables. Then

$$\lim_{\beta \to 0} \mathbb{E}\left(\frac{\|\mathbf{X} - \mathbf{Y}\|^\beta - 1}{\beta} \right) = \mathbb{E}\left(\log(\|\mathbf{X} - \mathbf{Y}\|) \right).$$

Proof. As already shown, the sequence

$$g_\beta(x) = x - 1 - \frac{x^\beta - 1}{\beta}$$

is non-negative for $\beta \in (0,1)$ and monotone in β. By the theorem of monotone convergence it follows that

$$\lim_{\beta \to 0} \mathbb{E}\left(\|\mathbf{X}-\mathbf{Y}\|\| - 1 - \frac{\|\mathbf{X}-\mathbf{Y}\|^{\beta}-1}{\beta} \right) = \mathbb{E}\left(\|\mathbf{X}-\mathbf{Y}\| - 1 - \log(\|\mathbf{X}-\mathbf{Y}\|) \right).$$

The assertion follows by the linearity of the expectation. □

In the following, we want to find conditions for the distributions of $\mathbf{X}$ and $\mathbf{Y}$ such that the expectation $\mathbb{E}\left(\log(\|\mathbf{X}-\mathbf{Y}\|)\right)$ is finite. Firstly, we can state the following remark.

Remark 9.1.4. Note that for discrete distributed d-dimensional random variables $\mathbf{X}$ and $\mathbf{Y}$ that take values on the same set $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ with corresponding probabilities $p_i = \mathbb{P}(\mathbf{X} = \mathbf{x}_i)$ and $q_i = \mathbb{P}(\mathbf{Y} = \mathbf{x}_i)$ it applies that

$$\mathbb{E}\left(\log(\|\mathbf{X}-\mathbf{Y}\|)\right) = \sum_{i,j} \log(\mathbf{x}_i - \mathbf{x}_j) p_i q_j = -\infty.$$

Therefore, let the distributions of $\mathbf{X}$ and $\mathbf{Y}$ be continuous. Then we can formulate the following two lemmas that set up further conditions on the distributions of $\mathbf{X}$ and $\mathbf{Y}$, such that $\mathbb{E}\left(\log(\|\mathbf{X}-\mathbf{Y}\|)\right) > -\infty$.

Lemma 9.1.5. If the density of $Z := \|\mathbf{X}-\mathbf{Y}\|$ is bounded on the interval $(0,1)$, it holds that $\mathbb{E}(\log(Z)) > -\infty$.

Proof. Let f_Z be the density of Z. Then

$$\begin{aligned} & \mathbb{E}(\log(\|\mathbf{X}-\mathbf{Y}\|) = \mathbb{E}(\log(Z)) \\ = & \int_0^{\infty} \log(z) f_Z(z) \mathrm{d}z \\ = & \int_0^{1} \log(z) f_Z(z) \mathrm{d}z + \int_1^{\infty} \log(z) f_Z(z) \mathrm{d}z. \end{aligned}$$

It is clear that the second integral is greater than $-\infty$, as the logarithm takes only positive values on the interval $(1,\infty)$.
Further, it holds for the first integral that

$$\int_0^1 \log(z) f_Z(z) \mathrm{d}z \geq \int_0^1 \log(z) \cdot b \, \mathrm{d}z = -b, \tag{9.3}$$

as the density f_Z is bounded by b on the interval $(0,1)$. Hence, the assertion follows. □

The following lemma imposes conditions on the distributions of $\mathbf{X}$ and $\mathbf{Y}$ that ensure that the density of Z is bounded.

Lemma 9.1.6. Let $\mathbf{X}, \mathbf{Y}$ be d-dimensional continuous distributed independent random variables. Further let the density of $\mathbf{X}$ be bounded. Then the density of $Z := \|\mathbf{X}-\mathbf{Y}\|$ is bounded on $(0,1)$.

Proof. If the density of $\mathbf{X}$ is bounded, it holds that $P(\mathbf{X} \in A) \leq b \cdot \lambda(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$. It follows that for all $\epsilon > 0$

$$\begin{aligned}
&P(a < Z < a + \epsilon) = P(a < \|\mathbf{X} - \mathbf{Y}\| < a + \epsilon) \\
&= \int P(a < \|\mathbf{X} - \mathbf{y}\| < a + \epsilon) P_{\mathbf{Y}}(\mathrm{d}\mathbf{y}) \\
&\leq \int b \cdot \lambda(A_\epsilon) P_{\mathbf{Y}}(\mathrm{d}\mathbf{y}) \\
&= b \cdot \lambda(A_\epsilon) = b \cdot \left(\frac{\pi^{\frac{d}{2}} (a+\epsilon)^d}{\Gamma\left(1 + \frac{d}{2}\right)} - \frac{\pi^{\frac{d}{2}} a^d}{\Gamma\left(1 + \frac{d}{2}\right)} \right) \\
&= b \cdot \frac{\pi^{\frac{d}{2}} \left((a+\epsilon)^d - a^d \right)}{\Gamma\left(1 + \frac{d}{2}\right)}.
\end{aligned}$$

As an illustration note that in the bivariate case A_ϵ is a circular ring. It follows that

$$\begin{aligned}
f_Z(a) &= \lim_{\epsilon \to 0} \frac{F_Z(a+\epsilon) - F_Z(a)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{P(a < Z < a + \epsilon)}{\epsilon} \\
&\leq \lim_{\epsilon \to 0} b \cdot \frac{\pi^{\frac{d}{2}} \left((a+\epsilon)^d - a^d \right)}{\epsilon \cdot \Gamma\left(1 + \frac{d}{2}\right)} \\
&= \lim_{\epsilon \to 0} b \cdot \frac{\pi^{\frac{d}{2}} \sum_{k=1}^{d} \binom{d}{k} a^{d-k} \epsilon^k}{\epsilon \cdot \Gamma\left(1 + \frac{d}{2}\right)} \\
&= \lim_{\epsilon \to 0} b \cdot \frac{\pi^{\frac{d}{2}} \sum_{k=1}^{d} \binom{d}{k} a^{d-k} \epsilon^{k-1}}{\Gamma\left(1 + \frac{d}{2}\right)} \\
&= b \cdot \frac{\pi^{\frac{d}{2}} d \cdot a^{d-1}}{\Gamma\left(1 + \frac{d}{2}\right)}.
\end{aligned}$$

As the bound

$$b \cdot \frac{\pi^{\frac{d}{2}} d \cdot a^{d-1}}{\Gamma\left(1 + \frac{d}{2}\right)}$$

is monotone increasing in a, this yields that f_Z is bounded by

$$b \cdot \frac{\pi^{\frac{d}{2}} d}{\Gamma\left(1 + \frac{d}{2}\right)}$$

on the interval $(0, 1)$. □

Example 9.1.7. Let X and Y be uniformly distributed on $(0, 1)$. Then we can define the random variable $Z := |X - Y|$. The corresponding density is $f_Z(t) = 2 - 2t$ for

$t \in [0, 1]$.
It follows that

$$\begin{aligned}&\mathbb{E}\left(\log(\|X - Y\|)\right) = \mathbb{E}\left(\log(Z)\right)\\ &= \int_0^1 \log(t)(2 - 2t)\mathrm{d}t = -3/2.\end{aligned}$$

Example 9.1.8. Let X and Y be distributed with density

$$f(x) = \begin{cases} 0, & x < 0 \\ 3x^2, & 0 \leq x \leq 1 \\ 0, & x > 1. \end{cases}$$

The corresponding distribution function is given by

$$F(x) = \begin{cases} 0, & x < 0 \\ x^3, & 0 \leq x \leq 1 \\ 1, & x > 1. \end{cases}$$

The distribution of $Z := |X - Y|$ can be calculated as

$$\begin{aligned}F_Z(x) &= \int (F(y + x) - F(y - x))f(y)\mathrm{d}y\\ &= \int F(y + x)f(y)\mathrm{d}y - \int F(y - x)f(y)\mathrm{d}y\\ &= \int_0^{1-x} (y + x)^3 \cdot 3y^2\mathrm{d}y + \int_{1-x}^1 3y^2\mathrm{d}y - \int_1^x (y - x)^3 \cdot 3y^2\mathrm{d}y\\ &= \frac{1}{20}\left(-2x^6 + 40x^3 - 90x^2 + 72x\right).\end{aligned}$$

Accordingly, we obtain the density

$$f_Z(x) = \frac{1}{20}(-12x^5 + 120x^2 - 180x + 72), \quad 0 \leq x \leq 1.$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left(\log(\|X - Y\|)\right) = \mathbb{E}(\log(Z))\\ &= \int_0^1 \log(z) \cdot \frac{1}{20}(-12x^5 + 120x^2 - 180x + 72)\mathrm{d}z\\ &= -2.\end{aligned}$$

Utilizing these results we can state the following theorem.

Theorem 9.1.9. Let $\mathbf{X}$ and $\mathbf{Y}$ be independent d-dimensional random variables with continuous distribution such that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and $\mathbb{E}\|\mathbf{Y}\|^2 < \infty$. Further, we assume that the density $f_{\mathbf{X}}$ of $\mathbf{X}$ and the density $g_{\mathbf{Y}}$ of $\mathbf{Y}$ is bounded. Then it holds that

$$\lim_{\beta \to 0} \frac{1}{\beta} \mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) \tag{9.4}$$

$$= 2 \cdot \mathbb{E}(\log \|\mathbf{X} - \mathbf{Y}\|) - \mathbb{E}(\log \|\mathbf{X} - \tilde{\mathbf{X}}\|) - \mathbb{E}(\log \|\mathbf{Y} - \tilde{\mathbf{Y}}\|). \tag{9.5}$$

Proof. Firstly, we proof that the expectations in (9.5) are finite. So let $\mathbf{X}$ and $\mathbf{Y}$ be independent d-dimensional random variables with continuous distributions. It holds that

$$\begin{aligned}
&\mathbb{E}\left(\log \|\mathbf{X} - \mathbf{Y}\|\right) = \mathbb{E}\left(\log\left(\sum_{i=1}^{n}(X_i - Y_i)^2\right)^{1/2}\right) \\
&= \mathbb{E}\left(\frac{1}{2}\log\left(\sum_{i=1}^{n}(X_i - Y_i)^2\right)\right) = \frac{1}{2}\mathbb{E}\left(\log\left(\sum_{i=1}^{n}(X_i - Y_i)^2\right)\right) \\
&\leq \frac{1}{2}\log\left(\mathbb{E}\left(\sum_{i=1}^{n}(X_i - Y_i)^2\right)\right) = \frac{1}{2}\log\left(\sum_{i=1}^{n}\mathbb{E}((X_i - Y_i)^2)\right),
\end{aligned}$$

where we used Jensen's inequality. So if $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and $\mathbb{E}\|\mathbf{Y}\|^2 < \infty$, it holds that $\mathbb{E}\left(\log \|\mathbf{X} - \mathbf{Y}\|\right) < \infty$. Further, if $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and $\mathbb{E}\|\mathbf{Y}\|^2 < \infty$, it obviously also holds that $1/\beta \cdot \mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y})$ is finite for all $\beta \in (0, 2)$.
According to Lemma 9.1.6 and Lemma 9.1.5, we have $\mathbb{E}\left(\log \|\mathbf{X} - \mathbf{Y}\|\right) > -\infty$ if the densities of $\mathbf{X}$ and $\mathbf{Y}$ are bounded. Analogously, this result follows for $\mathbb{E}(\log \|\mathbf{X} - \tilde{\mathbf{X}}\|)$ and $\mathbb{E}(\log \|\mathbf{Y} - \tilde{\mathbf{Y}}\|)$.
By the linearity of the expectation we can write

$$\frac{1}{\beta}\mathcal{E}^{(\beta)}(\mathbf{X}, \mathbf{Y}) = 2\mathbb{E}\left(\frac{\|\mathbf{X} - \mathbf{Y}\|^\beta - 1}{\beta}\right) - \mathbb{E}\left(\frac{\|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta - 1}{\beta}\right) - \mathbb{E}\left(\frac{\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^\beta - 1}{\beta}\right).$$

Thus, the assertion follows by Theorem 9.1.3. □

Analogously to the energy score and its associated divergence function, namely the energy distance, we can define a scoring rule corresponding to the function (9.5) which arises as a scaled limit of the energy distance. This scoring rule is given by

$$\mathrm{Sc}_{\log}(F, \mathbf{y}) = \mathbb{E}\left(\log(\|\mathbf{X} - \mathbf{y}\|)\right) - \frac{1}{2} \cdot \mathbb{E}\left(\log(\|\mathbf{X} - \tilde{\mathbf{X}}\|)\right).$$

In the following, we proof that this scoring rule is strictly proper with respect to a broad class of distributions. To answer this, we first recall the following important theorem which is the key result in the kernel construction of strictly proper scoring rules, see Appendix A.2.

Theorem 9.1.10 (Gneiting and Raftery (2007))**.** Let ψ be a continuous function on $[0, \infty)$ with ψ' completely monotone and not constant. For a Borel probability measure F on $\mathbb{R}^d$, let $\mathbf{X}$ and $\tilde{\mathbf{X}}$ be independent random vectors with distribution F. The scoring rule

$$S(F, \mathbf{y}) = \mathbb{E}_F \psi(\|\mathbf{X} - \mathbf{y}\|^2) - \frac{1}{2}\mathbb{E}_F \psi(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2)$$

is strictly proper relative to the class of the Borel probability measures F on $\mathbb{R}^d$ for which $\mathbb{E}_F \psi(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2)$ is finite.

Theorem 9.1.11. The scoring rule

$$\mathrm{Sc}_{\log}(F, \mathbf{y}) := \mathbb{E}\left(\log \|\mathbf{X} - \mathbf{y}\|\right) - \mathbb{E}\left(\log \|\mathbf{X} - \tilde{\mathbf{X}}\|\right) \tag{9.6}$$

is strictly proper relative to the class of continuous distributions such that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and the density $f_{\mathbf{X}}$ of $\mathbf{X}$ is bounded.

Proof. As $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ it follows that $\mathbb{E}\left(\log \|\mathbf{X} - \tilde{\mathbf{X}}\|\right) < \infty$ by Jensen's inequality, see for this the proof of Theorem 9.1.9.
According to Lemma 9.1.6 and Lemma 9.1.5, it holds that $\mathbb{E}\left(\log \|\mathbf{X} - \tilde{\mathbf{X}}\|\right) > -\infty$ if the distribution of $\mathbf{X}$ is continuous and the density of $\mathbf{X}$ is bounded. As the function $\psi(t) = \log(t^{1/2})$ is continuous on $[0, \infty)$ with $\psi'(t)$ completely monotone and not constant as it holds true that $\varphi(t) := \psi'(t) = 1/(2t)$ and

$$\varphi^{(k)} = (-1)^k \frac{k!}{2 \cdot t^{k+1}}.$$

Therefore, the assertion follows by Theorem 9.1.10. □

9.2. Some properties of the scoring rule

Clearly, the scoring rule $\mathrm{Sc}_{\log}$ can take both positive and negative values. That makes the interpretation of the score values received with $\mathrm{Sc}_{\log}$ harder when solely the score values are considered.
For instance, in case of the energy score it holds that the closer the score is to zero, the better is the quality of the forecast.

Example 9.2.1. So let for example the true distribution G be given by the uniform distribution on $(0, 1)$. Then the score of the perfect forecast is

$$\mathrm{Sc}_{\log}(G, G) = \frac{1}{2}\mathbb{E}(\log \|X - Y\|) = -\frac{3}{4},$$

where X and Y are independent and both distributed with distribution G.

Remark 9.2.2. When we assess the resulting score values with the Diebold-Mariano test, the scoring rule $\mathrm{Sc}_{\log}$ should perform just like the energy score with a very small parameter β as we received the scoring rule $\mathrm{Sc}_{\log}$ by a linear transformation with factor $1/\beta$ of the energy score. As the Diebold-Mariano test is invariant with respect to linear transformations the assertion follows.

Remark 9.2.3. The estimator can be implemented analogously to the energy score. Given the i.i.d. sample $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ as draws from the forecasting distribution F and the observation $\mathbf{y}$ from G, the first term can be estimated by

$$\widehat{\mathrm{ScD}_{\log}} := \frac{1}{M} \sum_{j=1}^{M} \log\left(\left|\left|\mathbf{X}^{(j)} - \mathbf{y}\right|\right|\right).$$

The second term has multiple plausible options for the estimation. The definition implies that we require the independent copy $\tilde{\mathbf{X}}$ of the forecasting distribution.
As the members of the ensemble $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ are i.i.d. we might use one half of this set as draws from $\mathbf{X}$ and the other half as draws from $\tilde{\mathbf{X}}$. So the resulting estimator is given by

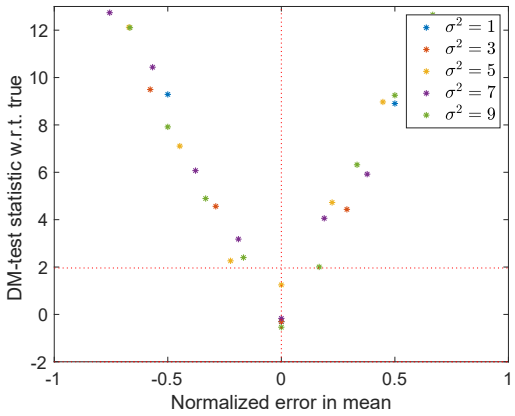
$$\widehat{\mathrm{ScI}^{iid}_{\log}} := \frac{1}{\lfloor 0.5M \rfloor} \sum_{j=1}^{\lfloor 0.5M \rfloor} \log\left(\left|\left|\mathbf{X}^{(j)} - \mathbf{X}^{(\lfloor 0.5M \rfloor + j)}\right|\right|\right).$$

Note that the sum only contains $\lfloor M/2 \rfloor$ summands, but they have nice statistical properties as they are i.i.d.
Another reasonable estimator is given by

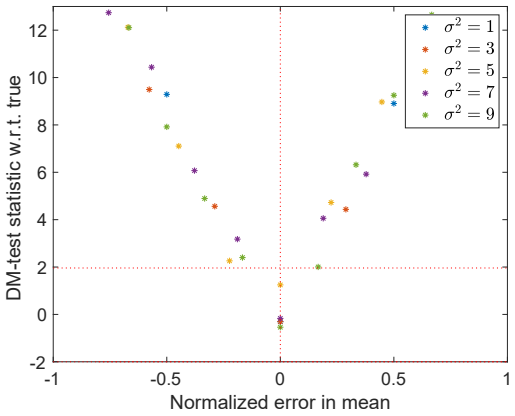
$$\widehat{\mathrm{ScI}^{band}_{\log}} := \frac{1}{M \cdot (M-1)} \sum_{j=1}^{M} \sum_{\substack{k=1 \\ k \neq j}}^{M} \log\left(\left|\left|\mathbf{X}^{(j)} - \mathbf{X}^{(k)}\right|\right|\right).$$

9.3. Simulation study: bivariate Gaussian process

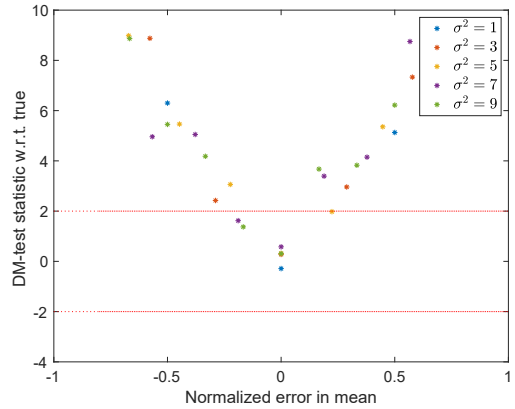
In the following, we repeat the simulation study of Pinson and Tastu (2013) for the newly defined scoring rule $\mathrm{Sc}_{\log}$ to study its discrimination ability.
We also compare it to the energy score with $\beta = 0.01$ and $\beta = 1$. It should hold that $\mathrm{Sc}_{\log}$ and $\mathrm{ES}_{0.01}$ perform quite similarly with respect to the discrimination ability assessed with the Diebold-Mariano test.
We first consider errors in mean, that is the real distribution Y is given by a bivariate Gaussian process with mean set to $\mu = 5$, correlation $\rho = 0.5$ and variance σ^2, where $\sigma^2 \in \{1, 3, 5, 7, 9\}$. In order to characterize the sensitivity to the process variance the following assessment is performed for different values of σ^2.
Further, let $\hat{\mu}$ be the mean parameter of the predictive distribution. We choose $\hat{\mu}$ out of the grid $\{0, 0.5, 1, 1.5, \ldots, 9.5, 10\}$. The DM-test statistic is evaluated as a function of the normalized error in mean $(\mu - \hat{\mu})/\sigma$.



(a) $\mathrm{Sc}_{\log}$



(b) Energy score with $\beta = 0.01$

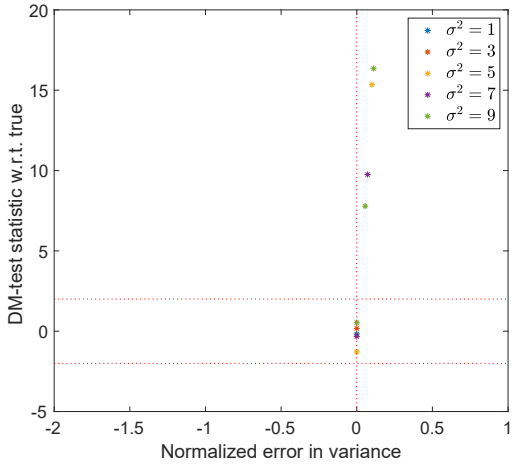


(c) Energy score with $\beta = 1$

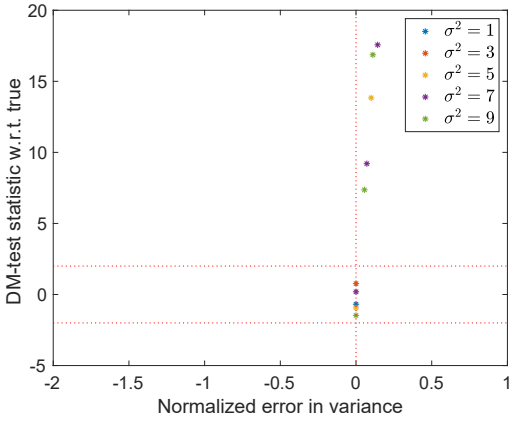
Figure 9.2.: DM-test statistic values for errors in mean, the statistics are capped to improve interpretability.

In case of a misspecified mean, the energy score with parameter $\beta = 0.01$ and $\mathrm{Sc}_{\log}$ perform almost exactly the same. The absolute values of the DM-test statistic for both scoring rules are only smaller than $1,96$ for the perfect forecasts. That is, all misspecified forecasting distributions are detected as not matching the true distribution. However, as already stated above this result was to be expected due to the invariance of the Diebold-Mariano test with respect to linear transformations.
Both scoring rules perform noticeably better than the energy score with $\beta = 1$ which falsely does not reject the null hypothesis in two cases.
Next we consider errors in variance. As in Section 7.3.2, we keep the mean and correlation parameters fixed as $\mu_{\mathbf{Y}} = (0, 0)'$ and $\rho = 0.5$ and evaluate $\Delta\mathrm{Sc}$ as a function of the relative prediction error in variance, which is defined as $(\sigma^2 - \hat{\sigma}^2)/\sigma^2$, where $\hat{\sigma}^2$ is the predictive variance. We choose $\hat{\sigma}^2$ out of the grid $\{0, 0.5, 1, \ldots, 9, 5, 10\}$. The assessment is performed for a set of σ^2 with $\sigma^2 \in \{1, 3, 5, 9\}$ to characterize the sensitivity to the process variance. Here all three scoring rules perform the same and do not

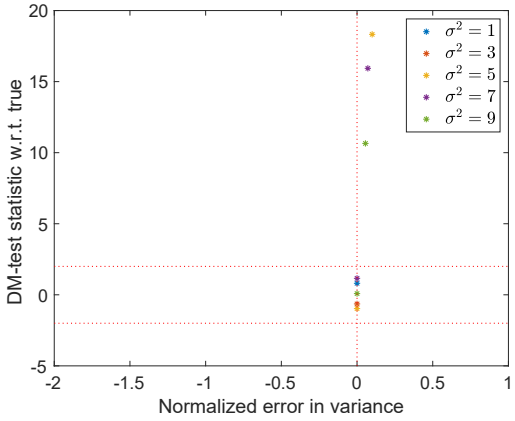
reject the null hypothesis only if the process variance is predicted correctly, i.e. only if the normalized error in variance is zero.



(a) $\mathrm{Sc}_{\log}$



(b) Energy score with $\beta = 0.01$



(c) Energy score with $\beta = 1$

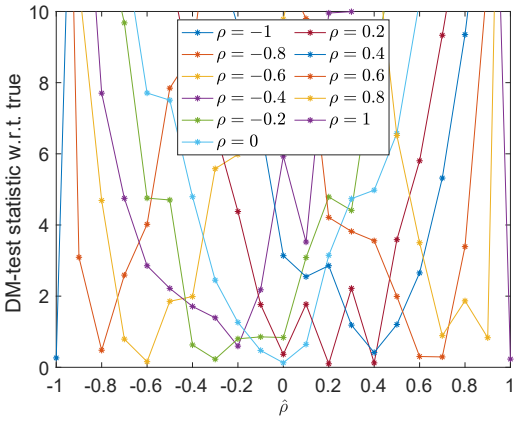
Figure 9.3.: DM-test statistic values for errors in variance, the statistics are capped to improve interpretability.

Lastly, we consider the discrimination ability of the scoring rule $\mathrm{Sc}_{\log}$ with respect to errors in correlation compared to the energy score with parameters $\beta = 0.01$ and $\beta = 1$. As above we keep the mean and variance parameters fixed as $\mu = (0, 0)$ and $\sigma^2 = 1$ and calculate the Diebold-Mariano test statistic values as a function of the predicted correlation $\hat{\rho}$. The assessment is performed for different values of

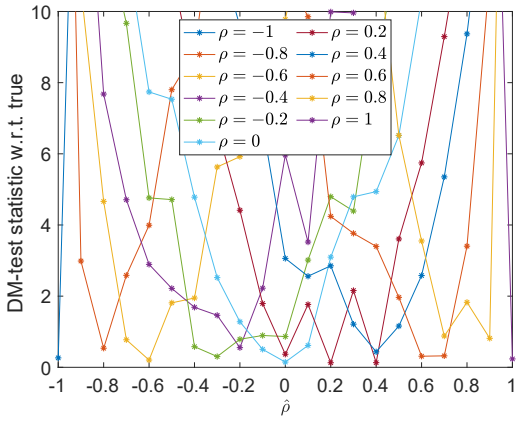
$$\rho \in \{-1, -0.8, \ldots, 0.8, 1\}.$$

For the predicted correlation we choose the denser grid

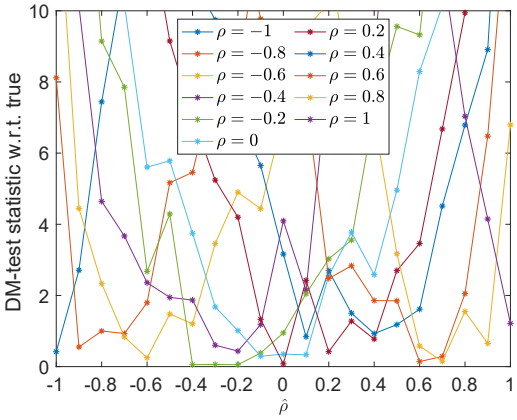
$$\hat{\rho} \in \{-1, -0.9, -0.8, \ldots, 0.9, 1\}.$$



(a) $\mathrm{Sc}_{\log}$



(b) Energy score with $\beta = 0.01$



(c) Energy score with $\beta = 1$

Figure 9.4.: DM-test statistic values for errors in correlation, the statistics are capped to improve interpretability.

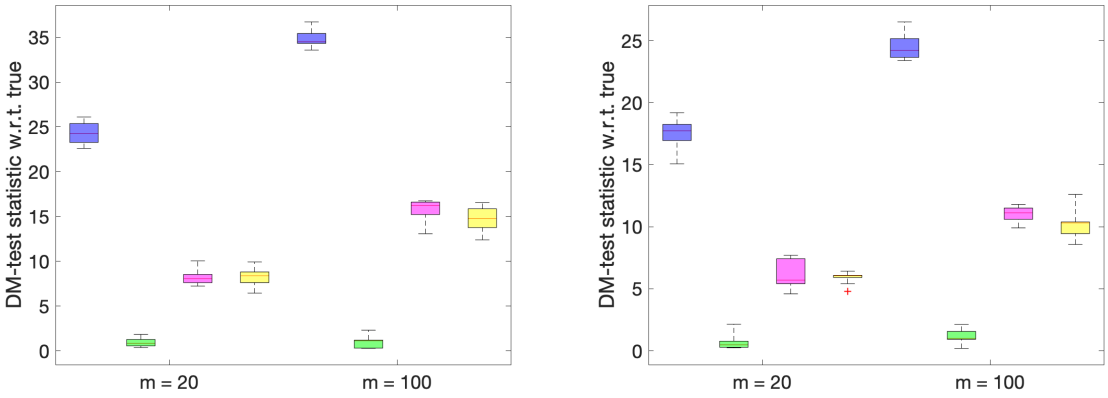
In all three simulation studies the scoring rule $\mathrm{Sc}_{\log}$ and the energy score with $\beta = 0.01$ admit roughly the same Diebold-Mariano test statistic values. Furthermore, both scoring rules admit a better discrimination ability than the energy score with the standard choice of $\beta = 1$.

9.4. Simulation study II revisited

We now repeat simulation study II, see Chapter 7.4, and assess the probabilistic forecasts with the newly defined scoring rule $\mathrm{Sc}_{\log}$. As the consideration of solely the score values is not meaningful, we constrict ourselves to the consideration of the values of the Diebold-Mariano test statistic.
Firstly, we consider the case of miscalibrated marginal distributions. So let the observations be realizations of a Gaussian distribution Y of dimension $d = 5$ with zero

mean, unit variance, and correlation structure given by the exponential model (7.2) with $r = 3$. The forecasting distributions X assign the true exponential correlation structure and have

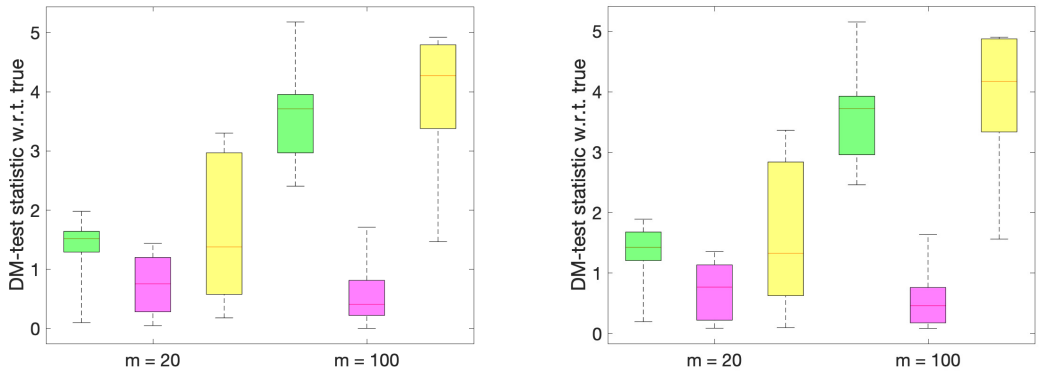
1. correct variances but biased means: $\mu_X = (-0.5, -0.25, 0, 0.25, 0.5)'$,
2. correct means and variances,
3. correct means but too large variances: $\sigma_X^2 = 1.5$,
4. correct means but too small variances: $\sigma_X^2 = 0.6667$.



(a) $\mathrm{Sc}_{\log}$ (b) Energy score with $\beta = 0.001$

Figure 9.5.: DM-test statistic values corresponding to the score values of the mean-biased, correct, over-, and underdispersive forecasts (blue, green, magenta, and yellow boxes) for ensembles sizes $m = 20$ and $m = 100$.

Afterwards, we consider errors in correlation strength for the dimension size $d = 15$. The true distribution is given by a zero mean, unit variance AR(1)-process with correlation function given in (7.2) with $r = 3$. The forecasting distributions have the same correlation model, but with $r = 2$, $r = 3$ and $r = 4.5$.



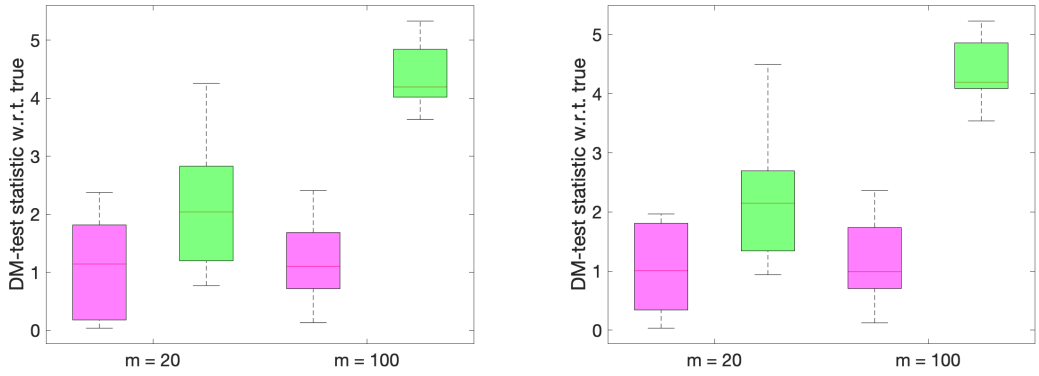
(a) $\mathrm{Sc}_{\log}$ (b) Energy score with $\beta = 0.001$

Figure 9.6.: DM-test statistic values corresponding to the score values of forecasts with too weak (green), accurate (magenta), and too strong (yellow) correlations compared to the observations.

Lastly, we consider the case of misspecified correlation models. Here we not just vary the correlation strength but the entire correlation model, see Scetion 7.4.3. We consider observations with zero mean, unit variance, and correlation function

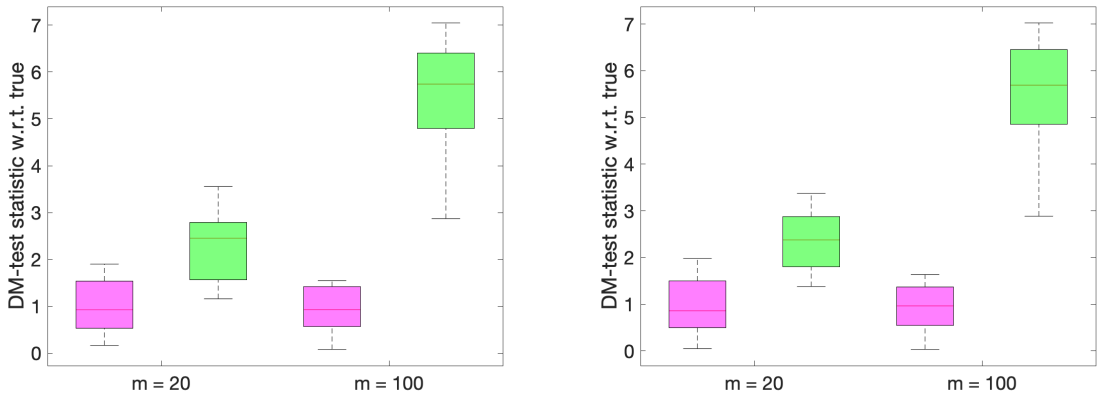
(i) $\mathrm{corr}(Y_i, Y_j) = \left(1 + \frac{|i-j|}{3}\right)^{-1}$, and

(ii) $\mathrm{corr}(Y_i, Y_j) = \exp\left(-\frac{|i-j|}{4}\right)\left[0.75 + 0.25 \cdot \cos\left(\frac{|i-j|\pi}{2}\right)\right]$.



(a) $\mathrm{Sc}_{\log}$ (b) Energy score with $\beta = 0.001$

Figure 9.7.: DM-test statistic values corresponding to the score values of forecasts with correct (left magenta boxplots), and incorrect (right green boxplots) correlation structure, where the correct correlation function is that using model (i) in Section 7.4.3, and the incorrect correlation function is the exponential model in (7.2) with $r = 3$.



(a) $\mathrm{Sc}_{\log}$ (b) Energy score with $\beta = 0.001$

Figure 9.8.: DM-test statistic values corresponding to the score values of forecasts with correct (left magenta boxplots), and incorrect (right green boxplots) correlation structure, where the correct correlation function is that using model (ii) in Section 7.4.3, and the incorrect correlation function is the exponential model in (7.2) with $r = 3$.

In all simulation studies the energy score with $\beta = 0.001$ and $\mathrm{Sc}_{\log}$ yield almost exactly the same DM-test statistic values which confirms the statement above that the scoring rule $\mathrm{Sc}_{\log}$ is the scaled limiting case of the energy score as β tends to zero.

10. Summary and conclusion

We analyzed the existing multivariate scoring rules in detail and also proposed a new scoring rule which results as a limiting case of the energy score as β tends to zero. In several simulation studies we found that the energy score, particularly, when the parameter β is small, discriminates well between forecasting distributions not only with different mean and variance parameters, but also among forecasting distributions with different interdependence structure between the components.
This contradicts the most common literature which states the energy score has a poor discrimination ability with respect to differences in the correlation structure between different forecasting distributions, see Pinson and Tastu (2013).
A crucial element in assessing the predictive performance of a probabilistic forecast or the discrimination ability of a scoring rule is the Diebold-Mariano test.
It can also be stated that the energy score is reasonably applicable to ensemble forecasting systems of smaller ensemble sizes $m = 100$ and even $m = 20$ which frequently arise in application, for instance, in weather and climate prediction due to the complexity of the underlying models and limited computation power.
The probably most important property of the energy score is that it is strictly proper to a broad class of distributions. Remember that the energy score is strictly proper relative to the class of Borel-probability distributions F such that $\mathbb{E}_F\|\mathbf{X}\|^\beta < \infty$. For a small parameter β this also includes the case of heavy-tailed distributions.
In contrast, the variogram score proposed by Scheuerer and Hamill (2015) is only proper, but not strictly proper and the Dawid-Sebastiani score is only strictly proper for Gaussian distributions. Therefore, these scoring rules are in general not able to identify the true distribution. The variogram score of order p, for instance, is not able to discriminate between distributions that have the same bias for every component. Moreover, large-scale errors that are the same for every component cancel out, see Section 7.3.1.
However, in some special cases the variogram score has a better discrimination ability between forecasting distributions that differ in their interdependence structure.
Therefore, the energy score with a small parameter β should be the scoring rule of choice in application.
The newly defined scoring rule $\mathrm{Sc}_{\log}$ that arises as a limiting case of the energy score is also strictly proper to a broad class of distributions, namely the continuous distributions with bounded densities and $\mathbb{E}\|\mathbf{X}\|^2 < \infty$. Similar to the energy score it is also readily applicable to ensemble forecasts.
Overall a general guideline for forecasters is to do ensemble forecasts with a huge sample size and evaluate on the full dimensional with either the energy score with a small

parameter β or the scoring rule $\mathrm{Sc}_{\log}$. Afterwards, it is crucial to assess the resulting score values with the Diebold-Mariano test for significance.
Optional the results can be backed up by other scoring rules. For instance, the variogram score can be utilized to additionally assess the interdependence structure of a forecasting distribution or the CRPS can be applied to the marginals.

An interesting question, that should be studied further, is for what reason the energy score with a smaller parameter β leads to a better discrimination ability. Apparantly, the function $x \mapsto |x|^\beta$ is strictly concave on $(0, \infty)$ for $0 < \beta < 1$ similar to the function $x \mapsto \log(|x|)$ utilized in the scoring rule $\mathrm{Sc}_{\log}$.
Furthermore, other functions than just $x \mapsto |x|^\beta$ in the definition of the energy score can be considered. The energy score also works out for every function which is the negative logarithm of an infinitely divisible characteristic function. So, for instance, the functions $x \mapsto \log(1 + |x|^2)$ or $x \mapsto 1 - \exp(|x|^2/2)$ could be studied in this context.

A. Mathematical appendix

A.1. Copulas

This section gives a short overview of *copulas*, mainly following Joe (2014).

Definition A.1.1. A function $C : [0,1]^d \to [0,1]$ is called *copula* if there is a random vector $\mathbf{U} = (U_1, \ldots, U_d)$ with uniformly distributed marginals on $(0,1)$ with distribution function C.

Consider a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with distribution function F. According to the important result from Sklar we are able to consider the univariate marginals of each variable X_i .

Theorem A.1.2. For every d-dimensional distribution function F with marginal distributions $F_1, \ldots, F_n$ there exists a copula C, such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

holds for every $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The copula C is unique if the marginals $F_1, \ldots, F_d$ are continuous.

Next we consider an important class of copulas. *Archimedean copulas* are often applied in dependence modeling as it easy to generate random numbers.

Definition A.1.3. A copula with generator function $\psi : [0, \infty) \to [0,1]$ is called *Archimedean copula* if it admits the functional form

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \ldots, \psi^{-1}(u_d)).$$

The function ψ is the *generator* of the copula.
This construction yields a copula if $\psi(0) = 1$, $\lim_{t\to\infty} \psi(t) = 0$, and the function ψ is n times differentiable with alternating signs, i.e.

$$(-1)^k \psi^{(k)}(t) \geq 0 \quad \text{for all } t \geq 0, \quad k = 1, \ldots, n.$$

This applies in particular if ψ is the Laplace transformed of a non-negative random variable.

Example A.1.4. We now consider the generator functions of the one parametric Gumbel and the one parametric Clayton copula.

- The copula with generator

$$\psi(t) = \exp\left(-t^{1/\theta}\right), \quad , t \geq 0, \, \theta \geq 1$$

  is called Gumbel copula.

- The copula with generator

$$\psi(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right), \quad , t \geq 0, \, \theta \geq 1$$

  is called Clayton copula.

Remark A.1.5. For $\theta \to 1$ we obtain independence, whereas for $\theta \to \infty$ the Gumbel copula converges to the Fréchet-Hoeffding upper bound.
The Gumbel copula which we consider in our simulation study is frequently used to model data with assymmetric dependence. This copula is well known for its ability to capture strong upper tail dependence and weak lower tail dependence. Due to the restriction of the dependence parameter θ the Gumbel copula can not reach the Fréchet-Hoeffding lower bound. This suggests that the Gumbel copula can not account for negative dependence.

An important characteristic of random vectors (X_1, X_2) are the *tail dependence coefficients* for describing dependencies in the tails. These coefficients give an indication of how likely it is that both random variables take very large/small values at the same time.

Definition A.1.6. The *upper tail dependence* λ_U and the *lower tail dependence* λ_L of a random vector (X_1, X_2) with continuous marginals are defined as

$$\lambda_L = \lim_{t \to 0} \mathbb{P}\left(X_2 \leq F_2^{-1}(t) | X_1 \leq F_1^{-1}(t)\right)$$

and

$$\lambda_U = \lim_{t \to 1} \mathbb{P}\left(X_2 > F_2^{-1}(t) | X_1 > F_1^{-1}(t)\right).$$

Remark A.1.7. In the definition, one can also swap the roles of X_1 and X_2 without changing the values. The upper and lower tail dependence depend only on the values of the copula on the diagonal.

Remark A.1.8. The parameter of the upper and lower tail dependence of the Gumbel copula can be calculated by $\lambda_U = 2 - 2^{1/\theta}$ and $\lambda_L = 0$.

Let us now consider the sampling a bivariate copula. In the case of continuous random variables we utilize the PIT to transform X_1 and X_2 to standard uniform random variables $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$. The copula is the joint distribution function of (U_1, U_2), i.e.

$$H(u_1, u_2) = \mathbb{P}(U_1 \leq u_1, U_2 \leq u_2), \quad 0 < u_1, u_2 < 1.$$

Simulations of bivariate copulas can be generated as follows. First we generate random numbers v_1 and v_2 from a uniform distribution on $(0, 1)$. Then, we set

$$\begin{aligned} u_1 &= v_1 \quad \text{and} \\ u_2 &= c_{u_1}^{-1}(v_2), \end{aligned}$$

where c_u^{-1} denotes the quasi inverse of the conditional distribution function

$$c_u(v) = \mathbb{P}(V \leq v | U \leq u) = \frac{\partial C(u.v)}{\partial u}.$$

The sampling of Archimedean copulas is considered in Hofert (2008).

A.2. Scoring rules and kernel functions

Now we turn to a far-reaching generalization of the energy score. In the course of this we firstly introduce the *kernel score*. The following statements in this section are all taken from Gneiting and Raftery (2007).

Let Ω be a nonempty set. A real-valued function g on $\Omega \times \Omega$ is said to be a negative definite kernel if it is symmetric in its arguments and $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j g(x_i, x_j) \leq 0$ for all positives integers n, all $a_1, \ldots, a_n \in \mathbb{R}$ that sum to 0, and all $x_1, \ldots, x_n \in \Omega$. In Berg, Christensen, and Ressel (1984) numerous examples of negatives definite kernels are given.

Theorem A.2.1. Let Ω be a Hausdorff space and let g be a non-negative, continuous negative definite kernel on $\Omega \times \Omega$. For a Borel probability measure P on Ω, let X and $\tilde{X}$ be independent random variables with distribution P. The scoring rule

$$\mathrm{Sc}(P, y) = \mathbb{E}_P g(X, y) - \frac{1}{2}\mathbb{E}_P g(X, \tilde{X}) \tag{A.1}$$

is proper relative to the class of the Borel probability measures P on Ω for which the expectation $\mathbb{E}_P g(X, \tilde{X})$ is finite.

Example A.2.2. If $\Omega = \mathbb{R}^d$, $\beta \in (0, 2)$ and $g(\mathbf{x}, \tilde{\mathbf{x}}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|^\beta$, where $\|\cdot\|$ denotes the Euclidean norm, then (A.1) recovers the energy score.

Utilizing this theorem the energy score can be generalized. For $\mathbf{x} \in \mathbb{R}^d$ and $\alpha \in (0, \infty]$, define the vector norm $\|\mathbf{x}\|_\alpha = (\sum_{i=1}^{d} |x_i|^\alpha)^{1/\alpha}$ if $\alpha \in (0, \infty)$ and $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq d} |x_i|$.
According to Schoenberg's theorem, see Berg, Christensen, and Ressel (1984), Theorem 3.2.2, and a strand of literature culminating in the work of Koldobsky (1992), and Zastavnyi (1993) it holds that if $\alpha \in (0, \infty]$ and $\beta > 0$, the kernel

$$g(\mathbf{x}, \tilde{\mathbf{x}}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|_\alpha^\beta, \quad \mathbf{x}, \tilde{\mathbf{x}} \in \mathbb{R}^d$$

is negative definite if and only if the following holds.

- $d = 1$, $\alpha \in (0, \infty]$, and $\beta \in (0, 2]$,
- $d \geq 2$, $\alpha \in (0, 2]$, and $\beta \in (0, \alpha]$,
- $d = 2$, $\alpha \in (2, \infty]$ and $\beta \in (0, 1]$.

Theorem A.2.1 can be sharpened in the crucial case of Euclidean sample spaces. Firstly recall that function ν on $(0, \infty)$ is said to be *completely monotone* if it has derivates $\nu^{(k)}$ of all orders and $(-1)^k \nu^{(k)}(t) \geq 0$ for all non-negative integers k and all $t > 0$.

Theorem A.2.3. Let ψ be a continuous function on $[0, \infty)$ with ψ' completely monotone and not constant. For a Borel probability measure F on $\mathbb{R}^d$, let $\mathbf{X}$ and $\tilde{\mathbf{X}}$ be independent random vectors with distribution F. The scoring rule

$$S(F, \mathbf{y}) = \mathbb{E}_F \psi(\|\mathbf{X} - \mathbf{y}\|^2) - \frac{1}{2}\mathbb{E}_F \psi(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2)$$

is strictly proper relative to the class of the Borel probability measures F on $\mathbb{R}^d$ for which $\mathbb{E}_F \psi(\|\mathbf{X} - \tilde{\mathbf{X}}\|^2)$ is finite.

The proof of this result follows directly by Mattner (1997). In particular, if $\psi(t) = t^{\beta/2}$, then the previous theorem ensures the strict propriety of the energy score relative to the class of the Borel probability measures P on $\mathbb{R}^d$ for which $\mathbb{E}_P \|\mathbf{X}\|_2^\beta$ is finite.

A.3. An upper bound for the discrimination ability of the energy score in the multivariate Gaussian case

Analogously to Pinson and Tastu (2013), we can calculate an upper bound of the relative change in score for the multivariate Gaussian case which depends on the dimension d and also the parameter β of the energy score in closed-form.
As an upper bound for the relative change in score we consider the case where the true underlying distribution G is given by a d-variate Gaussian distribution $Y \sim \mathcal{N}(0, \Sigma)$ with the same variance σ^2 on all dimensions, and a correlation of 1 between the different components, i.e.

$$\Sigma = \sigma^2 \mathbf{1}_{(d\times d)},$$

where $\mathbf{1}_{(d\times d)}$ is a $d \times d$-matrix of ones. Thus, a process observation is given by $\mathbf{y} = y\mathbf{1}_d$, where $\mathbf{1}_d$ is a d-dimensional vector of ones and y is a realization of$\mathcal{N}(0, \sigma^2)$.
The forecast F is given by the naive forecast, which totally neglects the interdependence structure, i.e. all components are independent. In this case the covariance matrix is given by $\hat{\Sigma} = \sigma^2 \text{diag}(\mathbf{1}_d)$. The mean of the naive forecast is assumed to be zero.

Firstly, we compute the expected energy score for the naive forecast. It holds that given a single process realization $\mathbf{y} = y\mathbf{1}_d$,

$$\|\mathbf{X} - \mathbf{y}\| = \sqrt{(X_1 - y)^2 + (X_2 - y)^2 + \dots (X_d - y)^2},$$

where

$$(X_i - y) \sim \mathcal{N}(-y, \sigma^2)$$

for all $i = 1, \dots, d$. Thus,

$$Z = \sum_{i=1}^{d} \frac{(X_i - y)^2}{\sigma^2} \sim \text{non} - \text{central Chi} - \text{squared}.$$

Consequently, following Harvey (1965), the parameters of the distribution are given by d and

$$\lambda = \frac{1}{2} \sum_{i=1}^{d} \frac{y^2}{\sigma^2}.$$

Furthermore, it holds that

$$\mathbb{E}_F \|\mathbf{X} - \mathbf{y}\|^\beta = \mathbb{E}_F \left(\sum_{i=1}^{d} (X_i - y)^2 \right)^{\beta/2} = \mathbb{E}_F \left(\frac{\sum_{i=1}^{d} (X_i - y)^2}{\sigma^2} \right)^{\beta/2} \cdot \sigma^\beta = \sigma^\beta \mathbb{E}_F(Z^{\beta/2}).$$

So we need the $\beta/2$-moment of the non-central Chi-squared distribution. Following Harvey (1965), it holds that

$$\begin{aligned}\mathbb{E}(Z^{\beta/2}) &= \sqrt{2}^{\beta}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\,{}_1\mathrm{F}_1\left(\frac{-\beta}{2},\frac{d}{2},-\lambda\right)\\ &= \sqrt{2}^{\beta}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\,{}_1\mathrm{F}_1\left(\frac{-\beta}{2},\frac{d}{2},-\frac{d}{2}\cdot\frac{y^2}{\sigma^2}\right),\end{aligned}$$

where ${}_1F_1$ denotes the confluent hypergeometric function. It follows that

$$\begin{aligned}&\mathbb{E}_G\left(\mathbb{E}_F\|\mathbf{X}-\mathbf{Y}\|^{\beta}\right)\\ &= \sigma^{\beta}\sqrt{2}^{\beta}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\int_{-\infty}^{\infty}{}_1\mathrm{F}_1\left(\frac{-\beta}{2},\frac{d}{2},-\frac{d}{2}\cdot\frac{y^2}{\sigma^2}\right)\cdot\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{y^2}{2\sigma^2}\right)\mathrm{d}y\\ &= \sigma^{\beta-1}\sqrt{2}^{\beta}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\int_{-\infty}^{\infty}{}_1\mathrm{F}_1\left(\frac{-\beta}{2},\frac{d}{2},-\frac{d}{2}\cdot\frac{y^2}{\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2\sigma^2}\right)\mathrm{d}y\\ &= \sigma^{\beta-1}\sqrt{2}^{\beta}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\int_{-\infty}^{\infty}{}_1\mathrm{F}_1\left(\frac{-\beta}{2},\frac{d}{2},-\frac{d}{2}\cdot\frac{y^2}{\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t}{2\sigma^2}\right)\sqrt{t}\,\mathrm{d}t.\end{aligned}$$

Utilizing (4) on page 822 of Gradshteyn and Ryzhik (2014), stating that

$$\int_0^{\infty}\exp(-st)t^{b-1}\,{}_1F_1(a,c,kt) = \Gamma(b)(s-k)^{-b}\,{}_2F_1(c-q,b;c;k/(k-s)$$

if $|s-k|>|k|$ and $\mathrm{Re}(b)>0, \mathrm{Re}(s)>\max(0,\mathrm{Re}(k))$. ${}_2F_1$ denotes the Gauss hypergeometric function in the above.
In our case we have $s=1/(2\sigma^2)$, $b=1/2$, $c=d/2$, $a=-\beta/2$ and $k=-n/(2\sigma^2)$. As

$$\begin{aligned}&|s-k| = \left|\frac{d+1}{2\sigma^2}\right| > \frac{d}{2\sigma^2} = |k|,\\ &\mathrm{Re}(b) = 1/2 > 0, \text{and}\\ &\mathrm{Re}(s) = \frac{1}{2\sigma^2} > 0 = \max(0,\mathrm{Re}(k))\end{aligned}$$

it follows that

$$\begin{aligned}&\mathbb{E}_G\left(\mathbb{E}_F\|\mathbf{X}-\mathbf{Y}\|^{\beta}\right)\\ &= (\sqrt{2}\sigma)^{\beta-1}\sqrt{2}^{\beta}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\frac{1}{\sqrt{\pi}}\Gamma(1/2)\left(\frac{d+1}{2\sigma^2}\right)^{-1/2}\,{}_2F_1\left(\frac{d+\beta}{2},\frac{1}{2};\frac{d}{2};\frac{d}{d+1}\right)\\ &= \sigma^{\beta}\sqrt{2}^{\beta}\frac{1}{\sqrt{d+1}}\frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\,{}_2F_1\left(\frac{d+\beta}{2},\frac{1}{2};\frac{d}{2};\frac{d}{d+1}\right).\end{aligned}$$

Next, we calculate $\mathbb{E}_G\left(\mathbb{E}_G\|\mathbf{X}-\tilde{\mathbf{X}}\|^{\beta}\right)$. Again we start with

$$\|\mathbf{X}-\tilde{\mathbf{X}}\| = \sqrt{(X_1-\tilde{X}_1)^2+(X_2-\tilde{X}_2)^2+\ldots(X_d-\tilde{X}_d)^2},$$

where $X_i, \tilde{X}_i \sim \mathcal{N}(0, \sigma^2)$ and mutually independent for all $i = 1, \ldots, d$. Let us define

$$Z = (\sum_{i=1}^{d} (X_i - \tilde{X}_i)^2)/(2\sigma^2).$$

As

$$\frac{X_i - \tilde{X}_i}{\sqrt{2}\sigma} \sim \mathcal{N}(0, 1),$$

Z follows a non-central Chi-squared distribution with parameters d and $\lambda = 0$. Therefore, it holds that

$$\begin{aligned} \mathbb{E}(Z^{\beta/2}) &= 2^{\beta/2} \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} {}_1F_1\left(-\beta/2, d/2, 0\right) \\ &= 2^{\beta/2} \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}. \end{aligned}$$

So altogether we have

$$\begin{aligned} \mathbb{E}_G \|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta &= \mathbb{E}_G \left(\sqrt{(X_1 - \tilde{X}_1)^2 + (X_2 - \tilde{X}_2)^2 + \ldots (X_d - \tilde{X}_d)^2}^{\beta} \right) \\ &= \mathbb{E}_G \left(\left(\frac{\sum_{i=1}^{d} (X_i - \tilde{X}_i)^2}{2\sigma^2} \right)^{\beta/2} \sqrt{2}^\beta \sigma^\beta \right) \\ &= \sigma^\beta \sqrt{2}^\beta \mathbb{E}_G(Z^{\beta/2}) \\ &= \sigma^\beta \sqrt{2}^\beta 2^{\beta/2} \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \\ &= \sigma^\beta 2^\beta \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}, \end{aligned}$$

and, thus, it also holds that

$$\mathbb{E}_G \left(\mathbb{E}_G \|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta \right) = \sigma^\beta 2^\beta \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

Altogether we receive the following closed-form solution for the expected score of the naive forecast

$$\mathrm{ES}_\beta(F, G) = \sigma^\beta \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \left[\frac{2^{\beta/2}}{\sqrt{d+1}} {}_2F_1\left(\frac{d+\beta}{2}, \frac{1}{2}; \frac{d}{2}; \frac{d}{d+1} \right) - 2^{\beta-1} \right].$$

Next, we consider the score of the perfect forecast. It holds that

$$\mathrm{ES}_\beta(G, G) = \mathbb{E}_G \left(\mathbb{E}_G \|\mathbf{X} - \mathbf{X}\|^\beta \right) - \frac{1}{2} \mathbb{E}_G \left(\mathbb{E}_G \|\mathbf{X} - \tilde{\mathbf{X}}\|^\beta \right) = \frac{1}{2} \mathbb{E}_G \left(\mathbb{E}_G \|\mathbf{X} - \mathbf{Y}\|^\beta \right),$$

where $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{Y}$ are i.i.d. with distribution G.
Note that

$$\mathbf{X} - \mathbf{Y} = (X_1 - Y_1, \ldots, X_d - Y_d).$$

As

$$\Sigma = \sigma^2 \mathbf{1}_{(d \times d)},$$

it holds that $X := X_1 = \cdots = X_d$ and $Y := Y_1 = \cdots = Y_d$. Therefore, we have

$$\|\mathbf{X} - \mathbf{Y}\| = \sqrt{d}\sqrt{(X-Y)^2}.$$

As $X \sim \mathcal{N}(0, \sigma^2)$, and given y fixed $X - y \sim \mathcal{N}(-1, \sigma^2)$ it follows that

$$\frac{X-y}{\sigma} \sim \mathcal{N}\left(-\frac{y}{\sigma}, 1\right),$$

thus,

$$Z := \left(\frac{X-y}{\sigma}\right)^2$$

follows a non-central Chi-squared distribution. With parameters $n = 1$ and $\lambda = \frac{1}{2}\frac{y}{\sigma^2}$ we have according to Harvey (1965)

$$\begin{aligned}
\mathbb{E}_G\left(\|\mathbf{X} - \mathbf{y}\|^\beta\right) &= \mathbb{E}_G\left[\left(\sqrt{d}\sqrt{(X-y)^2}\right)^\beta\right] \\
&= d^{\beta/2}\mathbb{E}_G\left[\left(\frac{(x-y)^2}{\sigma^2}\right)^{\beta/2}\sigma^\beta\right] \\
&= d^{\beta/2}\sigma^\beta\mathbb{E}\left[Z^{\beta/2}\right] \\
&= d^{\beta/2}\sigma^\beta 2^{\beta/2}\frac{\Gamma\left(\frac{1}{2}+\frac{\beta}{2}\right)}{\sqrt{\pi}}\,{}_1F_1\left(-\frac{\beta}{2}, \frac{1}{2}, -\frac{1}{2}\frac{y^2}{\sigma^2}\right).
\end{aligned}$$

As $Y \sim \mathcal{N}(0, \sigma^2)$,

$$\begin{aligned}
&\mathbb{E}_G\left(\mathbb{E}_G\|\mathbf{X} - \mathbf{Y}\|^\beta\right) \\
&= d^{\beta/2}\sigma^\beta 2^{\beta/2}\frac{\Gamma\left(\frac{1}{2}+\frac{\beta}{2}\right)}{\sqrt{\pi}}\int_{-\infty}^{\infty}{}_1F_1\left(-\frac{\beta}{2}, \frac{1}{2}, -\frac{1}{2}\frac{y^2}{\sigma^2}\right)\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{y^2}{2\sigma^2}\right)\mathrm{d}y \\
&= \sigma^{\beta-1}d^{\beta/2}2^{\beta/2-1/2}\pi^{-1}\Gamma\left(\frac{1}{2}+\frac{\beta}{2}\right)2\int_0^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right){}_1F_1\left(-\frac{\beta}{2}, \frac{1}{2}, -\frac{1}{2}\frac{y^2}{\sigma^2}\right)\mathrm{d}y \\
&= \sigma^{\beta-1}d^{\beta/2}2^{\beta/2-1/2}\pi^{-1}\Gamma\left(\frac{1}{2}+\frac{\beta}{2}\right)\int_0^{\infty}\exp\left(-\frac{t}{2\sigma^2}\right){}_1F_1\left(-\frac{\beta}{2}, \frac{1}{2}, -\frac{1}{2}\frac{t}{\sigma^2}\right)t^{-1/2}\mathrm{d}t.
\end{aligned}$$

Again, utilizing (4) on page 822 of Gradshteyn and Ryzhik (2014) stating that with

$$\begin{aligned}
a &= -\beta/2 \\
b &= 1/2 \\
k &= -\frac{1}{2\sigma^2} \\
s &= \frac{1}{2\sigma^2}
\end{aligned}$$

the conditions

$$\begin{aligned}
&|s-k| = \frac{1}{\sigma^2} > \frac{1}{2\sigma^2} = |k| \\
&\mathrm{Re}(b) = 1/2 > 0 \\
&\mathrm{Re}(s) = \frac{1}{2\sigma^2} > 0 = \max(0, \mathrm{Re}(k))
\end{aligned}$$

are fulfilled, we obtain

$$\begin{aligned}
&\mathbb{E}_G\left(\mathbb{E}_G\|\mathbf{X}-\mathbf{Y}\|^\beta\right) \\
&= \sigma^{\beta-1} d^{\beta/2} 2^{\beta/2-1/2} \pi^{-1} \Gamma\left(\frac{1}{2}+\frac{\beta}{2}\right) \Gamma(1/2)\sigma \, {}_2F_1\left(\frac{1+\beta}{2}, \frac{1}{2}, \frac{1}{2}; \frac{1}{2}\right) \\
&= 2^\beta d^{\beta/2} \sigma^\beta \pi^{-1/2} \Gamma\left(\frac{1+\beta}{2}\right).
\end{aligned}$$

Therefore, it holds that

$$\mathrm{ES}_\beta(G,G) = 2^{\beta-1} n^{\beta/2} \sigma^\beta \pi^{-1/2} \Gamma\left(\frac{1+\beta}{2}\right).$$

Consequently, for the relative change in score value we obtain

$$\begin{aligned}
&\Delta\mathrm{ES}_\beta(F) = \frac{\mathrm{ES}_\beta(F,G) - \mathrm{ES}_\beta(G,G)}{\mathrm{ES}_\beta(G,G)} \\
&= 2^{1-\beta} d^{-\beta/2} \sqrt{\pi} \frac{\Gamma\left(\frac{d+\beta}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{1+\beta}{2}\right)} \cdot \left(\frac{2^{\beta/2}}{\sqrt{d+1}} \, {}_2F_1\left(\frac{d+\beta}{2}, \frac{1}{2}, \frac{d}{2}; \frac{d}{d+1}\right) - 2^{\beta-1}\right) - 1.
\end{aligned}$$

B. Software

The introduced models in this thesis were implemented in MATLAB. The commercial software package is provided by MathWorks. We want to note that not all used functions are part of MATLAB's standard version. Some special toolboxes like the Econometrics toolbox are required. We used a student license provided by the university that has all the required toolboxes. In the following, the most important functions are described.

- **copularnd:** Generates copula random numbers. Possible copulas from the Archimedean copula family are Clayton, Frank and Gumbel.
- **random**: Generates random numbers from different kinds of distributions. Used distributions in this thesis are: Beta, Gamma, Gaussian and Uniform.

On demand the program code used in this thesis can be provided via sciebo.

List of Figures

List of Tables

Bibliography

Alexander, Carol et al. (2022). "Evaluating the discrimination ability of proper multivariate scoring rules". In: *Annals of Operations Research*, pp. 1–27.

Bailey, Andrew (2021). "Monetary Policy Report". In: URL: `https://www.bankofengland.co.uk/-/media/boe/files/monetary-policy-report/2022/february/monetary-policy-report-february-2022.pdf`.

Baringhaus, Ludwig and Franz, Carsten (2004). "On a new multivariate two-sample test". In: *Journal of multivariate analysis* 88.1, pp. 190–206.

– (2010). "Rigid motion invariant two-sample tests". In: *Statistica Sinica*, pp. 1333–1361.

Berg, Christian, Christensen, Jens Peter Reus, and Ressel, Paul (1984). *Harmonic analysis on semigroups: theory of positive definite and related functions*. Vol. 100. Springer.

Bianco, Bobby (2021). "What Does Probability of Precipitation Mean?" In: URL: `https://weatherworksinc.com/news/probability-of-precipitation-meaning`.

Bregman, Lev M (1967). "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR computational mathematics and mathematical physics* 7.3, pp. 200–217.

Brier, Glenn W (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Britton, Erik, Fisher, Paul, and Whitley, John (1998). "The Inflation Report projections: understanding the fan chart". In: *Chart* 8.10.

Bryc, W lodzimierz (1995). *Normal Distribution characterizations with applications*. Vol. 100.

Clements, Michael P (2005). "Density forecasts". In: *Evaluating Econometric Forecasts of Economic and Financial Variables*. Springer, pp. 103–123.

Cloke, HL and Pappenberger, Florian (2009). "Ensemble flood forecasting: A review". In: *Journal of hydrology* 375.3-4, pp. 613–626.

Corradi, Valentina and Swanson, Norman R (2006). "Bootstrap conditional distribution tests in the presence of dynamic misspecification". In: *Journal of Econometrics* 133.2, pp. 779–806.

Dawid, A Philip (1984). "Present position and potential developments: Some personal views statistical theory the prequential approach". In: *Journal of the Royal Statistical Society: Series A (General)* 147.2, pp. 278–290.

Diebold, Francis X (2015). "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests". In: *Journal of Business & Economic Statistics* 33.1, pp. 1–1.

Diebold, Francis X, Hahn, Jinyong, and Tay, Anthony S (1999). "Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange". In: *Review of Economics and Statistics* 81.4, pp. 661–673.

Diebold, Francis X and Mariano, Robert S (2002). "Comparing predictive accuracy". In: *Journal of Business & economic statistics* 20.1, pp. 134–144.

Diks, Cees, Panchenko, Valentyn, and Van Dijk, Dick (2011). "Likelihood-based scoring rules for comparing density forecasts in tails". In: *Journal of Econometrics* 163.2, pp. 215–230.

Ehm, Werner and Gneiting, Tilmann (2012). "Local proper scoring rules of order two". In: *The Annals of Statistics* 40.1, pp. 609–637.

Ehm, Werner, Gneiting, Tilmann, et al. (2016). "Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.3, pp. 505–562.

Frigyik, Béla A, Srivastava, Santosh, and Gupta, Maya R (2008). "Functional Bregman divergence and Bayesian estimation of distributions". In: *IEEE Transactions on Information Theory* 54.11, pp. 5130–5139.

Gel, Yulia, Raftery, Adrian E, and Gneiting, Tilmann (2004). "Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method". In: *Journal of the American Statistical Association* 99.467, pp. 575–583.

Gneiting, Tilmann (2011). "Making and evaluating point forecasts". In: *Journal of the American Statistical Association* 106.494, pp. 746–762.

Gneiting, Tilmann and Katzfuss, Matthias (2014). "Probabilistic forecasting". In: *Annual Review of Statistics and Its Application* 1, pp. 125–151.

Gneiting, Tilmann and Raftery, Adrian E (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477, pp. 359–378.

Gneiting, Tilmann, Raftery, Adrian E, et al. (2005). "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation". In: *Monthly Weather Review* 133.5, pp. 1098–1118.

Gneiting, Tilmann and Ranjan, Roopesh (2013). "Combining predictive distributions". In: *Electronic Journal of Statistics* 7, pp. 1747–1782.

Gneiting, Tilmann, Stanberry, Larissa I, et al. (2008). "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds". In: *Test* 17.2, pp. 211–235.

Good, Irving John (1952). "Probability and the weighing of evidence". In: *Rational Decisions, JRSS, Ser. B* 14, pp. 107–114.

– (1992). "Rational decisions". In: *Breakthroughs in statistics*. Springer, pp. 365–377.

Gradshteyn, Izrail Solomonovich and Ryzhik, Iosif Moiseevich (2014). *Table of integrals, series, and products*. Academic press.

Harvey, James Raymond (1965). *Fractional moments of a quadratic form in noncentral normal random variables*. North Carolina State University.

Hendrickson, Arlo D and Buehler, Robert J (1971). "Proper scores for probability forecasters". In: *The Annals of Mathematical Statistics* 42.6, pp. 1916–1921.

Hofert, Marius (2008). "Sampling archimedean copulas". In: *Computational Statistics & Data Analysis* 52.12, pp. 5163–5174.

Joe, Harry (2014). *Dependence modeling with copulas*. CRC press.

Jordan, Alexander, Krüger, Fabian, and Lerch, Sebastian (2017). "Evaluating probabilistic forecasts with scoringRules". In: *arXiv preprint arXiv:1709.04743*.

Jordan, T et al. (2011). "Operational Earthquake Forecasting: State of Knowledge and Guidelines for Implementation." In: *Annals of Geophysics*.

Knüppel, Malte (2015). "Evaluating the calibration of multi-step-ahead density forecasts using raw moments". In: *Journal of Business & Economic Statistics* 33.2, pp. 270–281.

Koldobsky, Alexander (1992). "Schoenberg's problem on positive definite functions". In: *arXiv preprint math/9210207*.

Koochali, Alireza et al. (2022). "Random Noise vs. State-of-the-Art Probabilistic Forecasting Methods: A Case Study on CRPS-Sum Discrimination Ability". In: *Applied Sciences* 12.10, p. 5104.

Kullback, Solomon and Leibler, Richard A (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Leutbecher, Martin and Palmer, Tim N (2008). "Ensemble forecasting". In: *Journal of computational physics* 227.7, pp. 3515–3539.

Mardia, KV, Kent, JT, and Bibby, JM (1979). "Multivariate analysis, 1979". In: *Probability and mathematical statistics. Academic Press Inc.*

Matheson, James E and Winkler, Robert L (1976). "Scoring rules for continuous probability distributions". In: *Management science* 22.10, pp. 1087–1096.

Mattner, Lutz (1997). "Strict definiteness of integrals via complete monotonicity of derivatives". In: *Transactions of the American Mathematical Society* 349.8, pp. 3321–3342.

Möller, Annette, Lenkoski, Alex, and Thorarinsdottir, Thordis L (2013). "Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas". In: *Quarterly Journal of the Royal Meteorological Society* 139.673, pp. 982–991.

Muller, Mervin E (1959). "A note on a method for generating points uniformly on n-dimensional spheres". In: *Communications of the ACM* 2.4, pp. 19–20.

Murphy, Allan H (1993). "What is a good forecast? An essay on the nature of goodness in weather forecasting". In: *Weather and forecasting* 8.2, pp. 281–293.

Murphy, Allan H and Winkler, Robert L (1987). "A general framework for forecast verification". In: *Monthly weather review* 115.7, pp. 1330–1338.

Ovcharov, Evgeni Y (2015). "Existence and uniqueness of proper scoring rules." In: *J. Mach. Learn. Res.* 16, pp. 2207–2230.

– (2018). *Proper scoring rules and Bregman divergence*.

Palmer, Tim N (2002). "The economic value of ensemble forecasts as a tool for risk assessment: From days to decades". In: *Quarterly Journal of the Royal Meteorologi-*

cal Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 128.581, pp. 747–774.

Parry, Matthew, Dawid, A Philip, and Lauritzen, Steffen (2012). "Proper local scoring rules". In: *The Annals of Statistics* 40.1, pp. 561–592.

Pinson, Pierre (2013). "Wind energy: Forecasting challenges for its operational management". In: *Statistical Science* 28.4, pp. 564–585.

Pinson, Pierre and Tastu, Julija (2013). *Discrimination ability of the energy score*. DTU Informatics.

Prudnikov, Anatoliui Platonovich et al. (1988). *Integrals and series*.

Rockafellar, R Tyrrell (1970). *Convex analysis*. Vol. 18. Princeton university press.

Salinas, David et al. (2019). "High-dimensional multivariate forecasting with low-rank gaussian copula processes". In: *Advances in neural information processing systems* 32.

Scheuerer, Michael and Hamill, Thomas M (2015). "Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities". In: *Monthly Weather Review* 143.4, pp. 1321–1334.

Stigler, Stephen M (1975). "The transition from point to distribution estimation". In: *Bull Int Stat Inst* 46, pp. 332–340.

Strähl, Christof and Ziegel, Johanna (2017). "Cross-calibration of probabilistic forecasts". In: *Electronic journal of statistics* 11.1, pp. 608–639.

Szekely, Gabor J and Rizzo, Maria L (2005). "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method". In: *Journal of classification* 22.2, pp. 151–184.

Székely, Gábor J (2003). "E-statistics: The energy of statistical samples". In: *Bowling Green State University, Department of Mathematics and Statistics Technical Report* 3.05, pp. 1–18.

Székely, Gábor J and Rizzo, Maria L (2013). "Energy statistics: A class of statistics based on distances". In: *Journal of statistical planning and inference* 143.8, pp. 1249–1272.

Taieb, Souhaib Ben et al. (2016). "Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression". In: *IEEE Transactions on Smart Grid* 7.5, pp. 2448–2455.

Tay, Anthony S and Wallis, Kenneth F (2000). "Density forecasting: a survey". In: *Journal of forecasting* 19.4, pp. 235–254.

Timmermann, Allan (2000). "Density forecasting in economics and finance". In: *Journal of Forecasting* 19.4, p. 231.

Uniejewski, Bartosz, Weron, Rafał, and Ziel, Florian (2017). In: *IEEE Transactions on Power Systems* 33.2, pp. 2219–2229.

Weron, Rafał and Ziel, Florian (2019). "Electricity price forecasting". In: *Routledge handbook of energy economics*. Routledge, pp. 506–521.

Winkler, Robert L and Murphy, Allan H (1968). ""Good" probability assessors". In: *Journal of Applied Meteorology and Climatology* 7.5, pp. 751–758.

Zastavnyi, VP (1993). "Positive definite functions depending on the norm". In: *Russian Journal of Mathematical Physics* 1, pp. 511–522.

Ziel, Florian and Berk, Kevin (2019). "Multivariate forecasting evaluation: On sensitive and strictly proper scoring rules". In: *arXiv preprint arXiv:1910.07325*.

Erklärung

Ich erkläre, dass mir die Promotionsordnung vom 6. August 2020 bekannt ist und von mir anerkannt wird.
Ich erkläre, dass ich weder früher noch gleichzeitig bei einer anderen Hochschule oder in einer anderen Fakultät ein Promotionsverfahren beantragt habe.
Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer, nicht angegebener Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Es wurden keine Dienste eines Promotionsvermittlers oder einer ähnlichen Organisation in Anspruch genommen.
Ich erkläre, dass zu den vorgeschlagenen Mitgliedern der Promotionskommission keine verwandtschaftlichen Beziehungen, keine Verwandtschaft ersten Grades, Ehe, Lebenspartnerschaft oder eheähnliche Gemeinschaft besteht.

Siegen, den 27. September 2022

Danksagung

Ich möchte mich an dieser Stelle bei Herrn Prof. Dr. Müller für die ausgezeichnete Betreuung bedanken. Prof. Dr. Oesting danke ich dafür, dass er sich bereit erklärt hat, die Arbeit zu begutachten. Ich danke meiner Familie für die vielfältige Unterstützung, die ich in den vergangenen Jahren durch sie erfahren habe.