Discrimination ability of the Energy score

Pinson, Pierre; Tastu, Julija

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Technical Report–2013–

Discrimination ability of the Energy score

Pierre Pinson[a]*, **Julija Tastu**[b]

[a] Department of Electrical Engineering, Technical University of Denmark
[b] Department of Applied Mathematics and Computer Science, Technical University of Denmark

Abstract

Research on generating and verification of multivariate probabilistic forecasts has gained increased interest over the last few years. Emphasis is placed here on the evaluation of forecast quality with the Energy score, which is based on a quadratic scoring rule. While this score may be seen as appealing since being proper, we show that its *discrimination* ability may be limited when focusing on the dependence structure of multivariate probabilistic forecasts. For the case of multivariate Gaussian process, a theoretical upper for such discrimination ability is derived and discussed. This limited discrimination ability may eventually get compromised by computational and sampling issues, as dimension increases.

Keywords: probabilistic forecasting, Energy score, discrimination, proper score, multivariate scenarios

1 Introduction

Probabilistic forecasting has gained increased attention over the last decade, both in terms of theoretical and of more practical developments. This phenomenon touches a wide range of applications, from economics and finance [1, 2], to earthquake prediction [3], while it also has wide appeal in meteorology [4], and for weather-related processes like renewable energy production [5, 6] and floods [7]. Such a focus on probabilistic forecasting is justified by the fact that, even if forecast users may prefer being provided with single-valued forecasts easier to handle in decision-making processes, those should be preferably extracted from probabilistic forecasts in a decision-theoretical framework, by accounting for user-specific loss functions (see, e.g., [8]).

Probabilistic forecasts optimally take the form of predictive densities for the stochastic process considered. If decisions to be made involve a univariate stochastic process only, or if they do not require modeling a dependence structure (multivariate or spatio-temporal), then only marginal predictive densities are required. These are referred to as marginal since being issued for each variable, location and lead time, individually. In the more general case of decision-making requiring to account for a dependence structure, probabilistic forecasts then ought to consist of multivariate predictive densities, hence describing both the marginal densities and the dependence structure.

* *Contact author:*
Technical University of Denmark
Department of Electrical Engineering
email: pp@dtu.dk

Evaluating probabilistic forecasts is more complex than evaluating single-valued predictions, even though some of the basic principles may be seen as similar. The main lines of probabilistic forecast verification frameworks (and underlying theoretical concepts) can be found in, e.g., [9], [10] and [11] among others. Such verification techniques may rely on scores, diagnostic tools, and possibly hypothesis testing. For the case of predictive densities, both univariate and multivariate, a number of scores have been proposed and discussed, see for instance [9] and [12]. Emphasis is placed here on quadratic scoring rules for multivariate predictive densities, that is, more precisely, on the Energy score introduced by [9]. Our aim is to discuss its discrimination ability, i.e., its ability to give significantly different score values to forecasts of different quality.

The manuscript is organized as follows. Section 2 recalls the background on probabilistic forecast verification based on scoring rules, while insisting on the fact that propriety of a score does not imply any discrimination ability. This section also illustrates how the Energy score has a substantially higher discrimination ability when misspecifying the mean of multivariate distributions, than in the case of misspecifying their variance or the dependence structure. Subsequently, some theoretical results are given in Section 3 giving a higher bound on differences in Energy score values for the case of misspecification of the dependence structure of multivariate predictive densities, also accounting for the dimension of these forecasts. Note that the discussion and results are produced for the Gaussian case only, though it is commonly used in practice today, for instance for short-term forecasts of surface wind speeds [13, 14] or for seasonal forecasts of sea-surface temperatures [15]. The necessary mathematical developments for obtaining these results are gathered in an Appendix at the end of the manuscript. Finally, Section 4 develops into a discussion on how to maximize the discrimination ability of quadratic scoring rules for multivariate probabilistic forecasts, also considering perspectives for future work on multivariate probabilistic forecast verification.

2 Discrimination ability of the Energy score

2.1 General setup

Let us place ourselves in a framework where a forecaster aims at issuing multivariate probabilistic forecasts in the form of predictive densities. He therefore considers a multivariate random variable $\mathbf{Y} \in \mathbb{R}^n$, $n > 1$. Write G the true distribution of $\mathbf{Y}$, $\mathbf{Y} \sim G$, while F is the multivariate predictive density issued by the forecaster at some point prior or equal to the current time. Time indices are not used here, since the results are independent of the time when the forecast is issued, and of the lead time considered. As an example, the multivariate random variable may be surface wind speed, expressed in its zonal and meridional components, as in the case of [13], [14], and [16]. More generally in meteorological prediction, it could also consist in a set of meteorological variables, e.g. wind speed, precipitation, etc., as in the case of [17]. In addition, the dependencies may not only be between various variables, but also for various geographical locations, and/or a set of times in the future [18]. Other setups exist in econometrics and finance related prediction problems, as for the example of the simultaneous confidence regions of [19] among others.

2.2 From propriety of scoring rules to their discrimination ability

When employing skill scores for probabilistic forecast verification, it is required that they are based on proper scoring rules, to ensure that forecasters really aim at issuing better forecasts, instead of focusing on hedging the score only. A scoring rule Sc is defined as a functional assigning a value to the association of a predictive density F with an observation $\mathbf{y}$ from the real density G of the random variable,

$$\mathrm{Sc} : (F, \mathbf{y}) \rightarrow \mathrm{Sc}(F, \mathbf{y}) \in \mathbb{R} \tag{1}$$

Formally, following the presentation by, e.g., [20], a scoring rule Sc (and associated score) for multivariate

predictive densities, is said to be proper if and only if

$$\mathrm{Sc}(G, \mathbf{y}) \leq \mathrm{Sc}(F, \mathbf{y}), \tag{2}$$

meaning, using simple words, that actual densities for the stochastic process are to be assigned the lowest possible score value. This result is for a negatively-oriented score, for which lowest values are seen as best. For simplicity, we only consider negatively-oriented scores in the following. Better, the scoring rule is strictly proper if only the actual densities get the lowest score value, i.e.,

$$\mathrm{Sc}(G, \mathbf{y}) < \mathrm{Sc}(F, \mathbf{y}). \tag{3}$$

Propriety is a property of scoring rules involving a predictive density and the real density of the stochastic process. In practice, that real density is not available anyway, and one is left with comparing alternative predictive densities, say, $F^{(1)}$ and $F^{(2)}$ generated by two rival forecasters. Propriety does not ensure that a difference in quality between $F^{(1)}$ and $F^{(2)}$ would yield a difference between $\mathrm{Sc}(F^{(1)}, \mathbf{y})$ and $\mathrm{Sc}(F^{(2)}, \mathbf{y})$, for any observation $\mathbf{y}$ drawn from G. Consequently, we refer to as *discrimination* the property of the scoring rule Sc such that

$$F^{(1)} \succ F^{(2)} \iff \mathrm{Sc}(F^{(1)}, \mathbf{y}) < \mathrm{Sc}(F^{(2)}, \mathbf{y}) \tag{4}$$

for any observation $\mathbf{y}$ drawn from G. In the above, $F^{(1)} \succ F^{(2)}$ means that $F^{(1)}$ genuinely is of higher quality than $F^{(2)}$. A scoring rule is then said to have a high discrimination ability if differences in quality between predictive densities are equivalent to significant differences in score values. At the opposite, a scoring rule is said to have no discrimination ability in the case where the same score values are assigned to predictive densities of different quality. One notes that proper score values may not need to have any discrimination ability, since possibly assigning the same score values to all predictive densities F, as well as G which is that for the actual random variable $\mathbf{Y}$. The situation is different for strictly proper scoring rules, since they should at least discriminate locally in the neighborhood of G. For densities F significantly different from G, however, there is no insurance that the score discriminate among predictive densities. It is to be noted that this concept of discrimination is inspired by the work of [21], who introduced some of the key concepts in forecast verification. Here, however, discrimination is a property of the score, not of the forecast themselves.

2.3 Characterizing the discrimination ability of the Energy score

Given the predictive density F and corresponding realization $\mathbf{y}$, the Energy score Es is defined as

$$\mathrm{Es}(F, \mathbf{y}) = \mathbb{E}_F\left[||\mathbf{X} - \mathbf{y}||\right] - \frac{1}{2}\mathbb{E}_F\left[||\mathbf{X} - \mathbf{X}'||\right], \tag{5}$$

where $\mathbf{X}$ and $\mathbf{X}'$ are independent random draws from F, while $||.||$ denotes the Euclidean norm. Computationally efficient estimators for the Energy scores were introduced in [13].

The corresponding expected Energy score, $\overline{\mathrm{Es}}$, can be calculated as the expectation of the Energy score in Eq. (5) over all potential observations of $\mathbf{Y}$, i.e.,

$$\overline{\mathrm{Es}}(F, G) = \mathbb{E}_G\left[\mathbb{E}_F\left[||\mathbf{X} - \mathbf{Y}||\right] - \frac{1}{2}\mathbb{E}_F\left[||\mathbf{X} - \mathbf{X}'||\right]\right]. \tag{6}$$

In order to analyze the discrimination ability of the Energy score, we define a metric to be used in the following, corresponding to relative changes in Energy score values, induced by differences between predicted and actual multivariate density of the stochastic process. Considering multivariate Gaussian processes, such differences may relate to prediction errors in the mean, variance, or interdependence structure. The relative change in expected Energy score is defined based of expected Energy score values for F and G,

$$\Delta\mathrm{Es} = \frac{\overline{\mathrm{Es}}(F, G) - \overline{\mathrm{Es}}^*}{\overline{\mathrm{Es}}^*}, \tag{7}$$

where the Energy score value $\overline{\mathrm{Es}}^* = \overline{\mathrm{Es}}(G, G)$ directly comes from the inherent uncertainty of the random variable $\mathbf{Y}$.

2.4 Illustrating the discrimination ability of the Energy score for multivariate Gaussian processes

To illustrate this concept of discrimination ability we make the following simulation study. We assume that a real process generating density is G, corresponding to a bivariate Gaussian. The process is simulated by considering 1000 instances: at each of these instances the process realization $\mathbf{y}$ is given by a single draw from G.

Suppose, there are two competing forecasters. One of them issues forecasts based on the real process generating density G — the perfect forecast. In parallel, the other forecaster issues alternative predictive density F. In order to compare the forecasting approaches, we estimate the corresponding Energy score values and compare them. For the calculation of these Energy score values, 1000 random draws from both densities are used, then employing the computationally efficient estimator proposed by [13]. Our main interest is to see how differences between G and F reflect in the corresponding Energy score values.

The process generating density, G, is bivariate Gaussian with a mean given by $\boldsymbol{\mu} = [\mu\ \mu]^\top$ and a covariance structure

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{8}$$

Then at every time step t a single process realization $\mathbf{y} = [y_1\ y_2]^\top$ of $\mathbf{Y}$ is such that

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{9}$$

The following differences between G and F have been considered:

- *Error in mean* corresponds to the case where the second forecaster makes an error in centering the predictive density only. In this case F is given by a bivariate Gaussian with a well specified covariance structure and a misspecified mean. That is, for every time step this forecaster issues a forecast describing the multivariate density for a random variable $\mathbf{X}$ such that

  $$\mathbf{X} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}), \tag{10}$$

  where $\hat{\boldsymbol{\mu}} = [\hat{\mu}\ \hat{\mu}]$. The resulting difference between F and G is depicted in Fig. 1(a).

- *Error in variance* corresponds to case where the forecaster makes an error while specifying the variance only. More specifically, for every time step the forecaster issues a forecast describing the multivariate density for a random variable $\mathbf{X}$ such that

  $$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}), \tag{11}$$

  where

  $$\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{12}$$

  The resulting difference between F and G is depicted in Fig. 1(b).

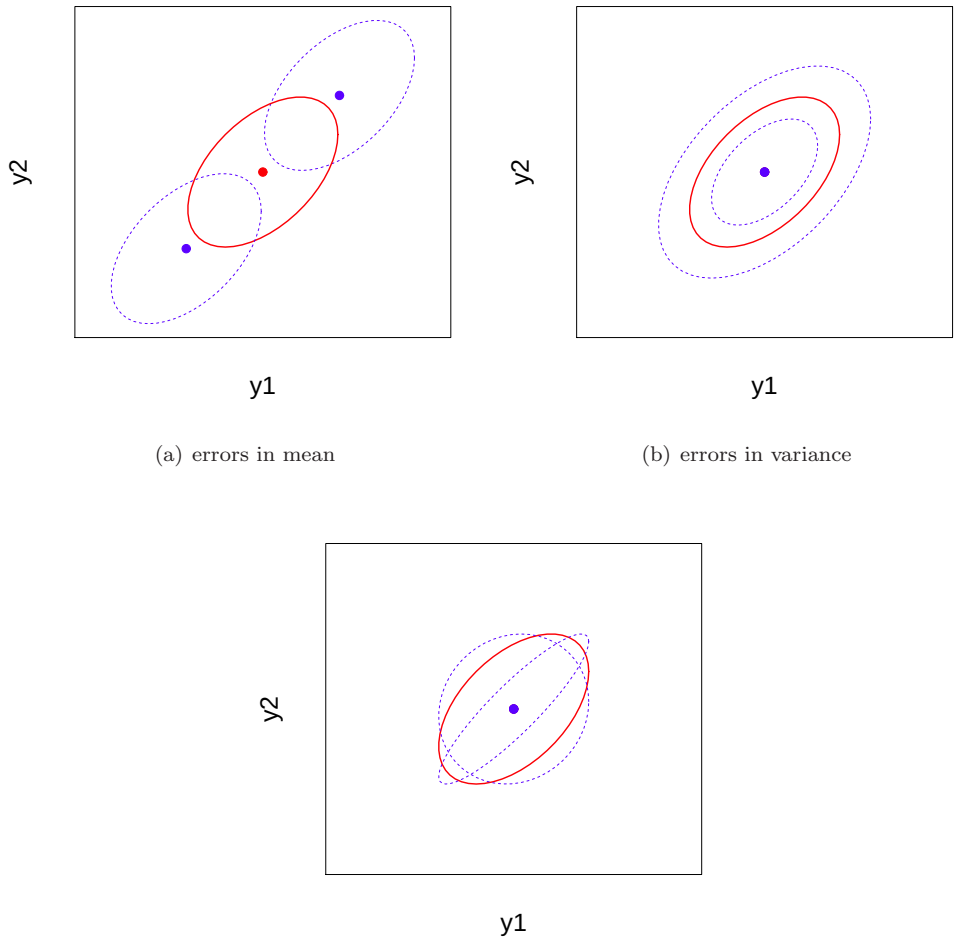
- *Error in correlation* corresponds to cases where the forecaster makes an error in describing the dependency structure, while well specifying the mean and the variance of the process. More specifically, for every time step the forecaster issues a forecast describing the multivariate density for a random variable $\mathbf{X}$ such that

  $$X \sim \mathcal{N}(\mu, \hat{\boldsymbol{\Sigma}}) \tag{13}$$

  where

  $$\hat{\boldsymbol{\Sigma}} = \sigma^2 \begin{bmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{bmatrix} \tag{14}$$

  The resulting difference between F and G is depicted in Fig. 1(c).



(a) errors in mean

(b) errors in variance

(c) errors in correlation

Figure 1: *Illustration of different misspecification in the process generating density. Each density is represented by a single iso-density contour. For every density the volume over the area bounded by the ellipse equals α. Here α denotes a pre-defined probability that a random draw from the corresponding density falls inside the ellipse. Red solid lines represents the real process generating density, while dotted blue lines correspond to predictive densities. Errors in mean are shown in (a). They correspond to predictive densities being shifted variants of the real process generating density. The shift along the major axis of the ellipse (45°) has been considered in the simulation work. Errors in variance are shown in (b). They correspond to an inflation ($\hat{\sigma^2} > \sigma^2$) or a deflation ($\hat{\sigma^2} < \sigma^2$) of the ellipse. Finally, errors in correlation are shown in (c). They correspond to stretching the ellipse in the direction of its major axis ($\hat{\rho} > \rho$) or its minor axis ($\hat{\rho} < \rho$)*

Let us first look at the discrimination ability of the Energy score for errors in mean. μ is set to $\mu = 5$, while the correlation value is fixed to $\rho = 0.5$ (any other values would yield qualitatively similar results). The relative change in Energy score ΔEs is evaluated as a function of the normalized error in predicting the mean parameter for $\mathbf{Y}$ (See Fig. 2(a)). That normalized prediction error is defined as $(\mu - \hat{\mu})/\sigma$. This assessment is performed for a number of values of σ^2 ($\sigma^2 \in \{1, 3, 5, 9\}$), in order to characterize the sensitivity to the process variance. A key result here is that, independently of the process variance, the effect of the relative prediction error for μ remains the same. Besides, since the Energy score is based on an Euclidean distance, the relative change in Energy score only depends on the magnitude of the normalized error in mean and not on its direction along the given translation axis (see Fig. 1(a)).

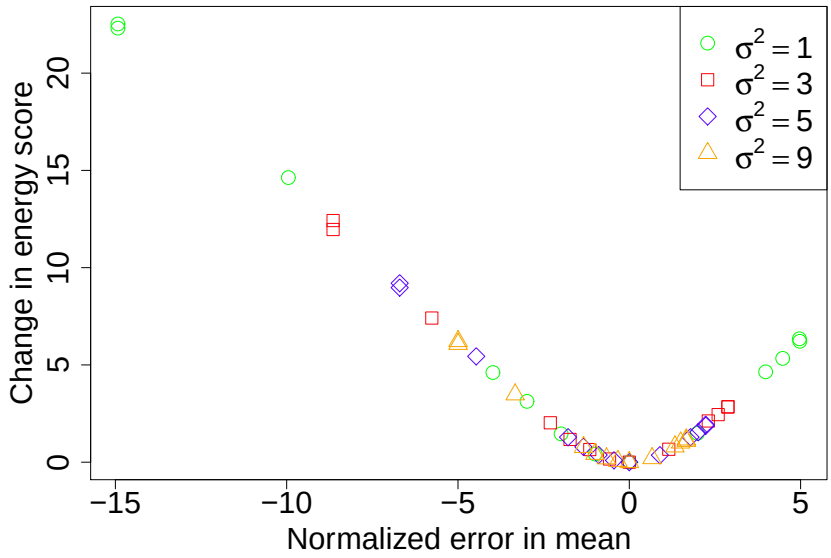
For the case of errors in variance, the setup is fairly similar, with mean and correlation parameters fixed to $\mu = 0$ and $\rho = 0.5$. A set of values for σ^2 are considered, i.e., $\sigma^2 \in \{1, 3, 5, 9\}$. Predictive densities F there only differ in terms of process variance, where the relative prediction error in variance is defined as $(\sigma^2 - \hat{\sigma}^2)/\sigma^2$. A plot of the relative change in Energy score ΔEs as a function of that relative prediction error in variance is depicted in Fig. 2(b). Here again, the relative change in Energy score follows similar patterns, independently of the actual process variance. The discrimination ability of the score is not symmetric, since the score increase for sharper densities is steeper than that for predictive densities that are too wide. Finally, comparing the discrimination ability of the Energy score for the mean and variance parameters, it is clear that the scale of variations in Figs. 2(a) and 2(b) are totally different (by a factor of 50), the score being clearly more sensitive to differences in the mean parameter than for the variance.

We finally look at errors in correlation. A plot for the relative change in Energy score as a function of predicted correlation is depicted in Fig. 3. The results were with $\sigma^2 = 1$ and $\mu = 0$. The results obtained for other values of σ^2 and μ were qualitatively similar. ΔEs describes how much the Energy score changes when instead of the real process generating density, the forecaster issues the predictive density F. The largest ΔEs is obtained when the real process generating density is perfectly correlated ($\rho = 1$), while the forecaster totally neglect the correlation by setting $\hat{\rho} = 0$. This is the extreme case for which, visually, $\Delta Es \approx 0.07$. A theoretical value for this upper bound may be derived analytically, as will be done in the following section. In practice this upper bound is seldom reached, since it corresponds to a very special case of a perfectly correlated bivariate Gaussian process. This means that for any realization $\mathbf{y}$ for G, $y_1 = y_2$. In such an extremes case, the only way for the forecaster to obtain an Energy score value (in expectation) of only 7% worse than if issuing perfect forecasts, is to totally ignore this strong dependency and assume independence of the components of $\mathbf{Y}$.

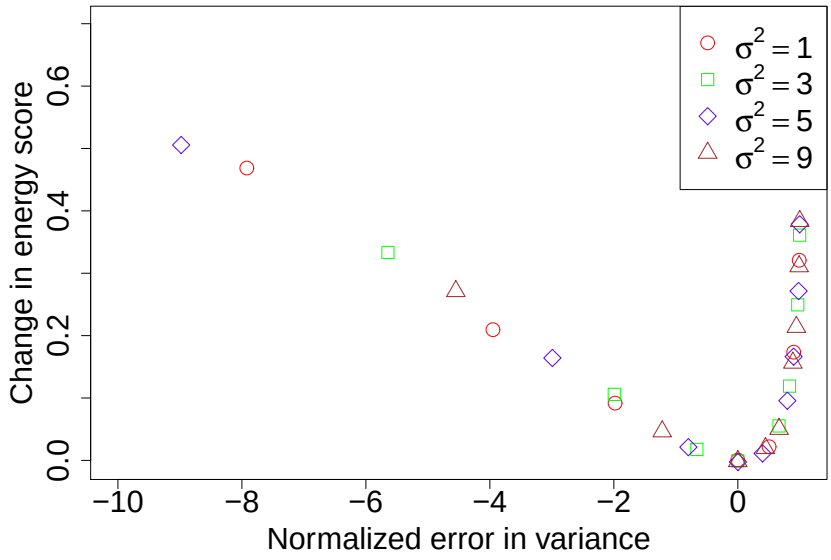
In practice, bivariate Gaussian processes are seldom perfectly correlated. One can notice that when the actual correlation, ρ, is less than 1, the maximum penalty (in expectation) stemming from errors in correlation becomes substantially less than 0.07. Already with $\rho = 0.8$, the maximum ΔEs is not even reaching 0.04. A conclusion from this plot is that the upper bound that may be derived by considering perfectly correlated generating densities, and predictive densities with independence of individual components, would be rarely met in most practical applications. For instance, if $\rho = 0.8$ and the forecaster tries predicts of a correlation $\hat{\rho} = 0.4$, then $\Delta Es < 0.02$ only. Another important factor is that the increase in ΔEs is steeper in the case for which $\hat{\rho} > \rho$. This also reduces the motivation of forecasters to move from the assumption of independence to try and capture the actual ρ.

3 Some theoretical results on the discrimination ability of the Energy score

In this section, emphasis is placed on our main result, which consists in an upper bound on the discrimination ability of the Energy score for multivariate Gaussian processes for the case of errors in correlation. Such a theoretical upper bound is of importance, since justifying the limited differences in Energy score values reported in various recent works focusing on multivariate probabilistic forecasts and the predictive modeling of interdependence structures, e.g. [13], [14], also giving some insight on results for Gaussian-copula based modeling of multivariate predictive densities as in [17]. Some other works, e.g., [16], appear to report results that go significantly beyond this theoretical upper bound, for reasons we cannot explain. Simulations studies performed with different variances did not show a significant change in this theoretical



(a) errors in mean



(b) errors in variance

Figure 2: *Discrimination ability of the Energy score assessed with ΔEs, in terms of its sensitivity to prediction errors in mean (a) and variance (b) for bivariate Gaussian predictive densities (hence for $n = 2$).*

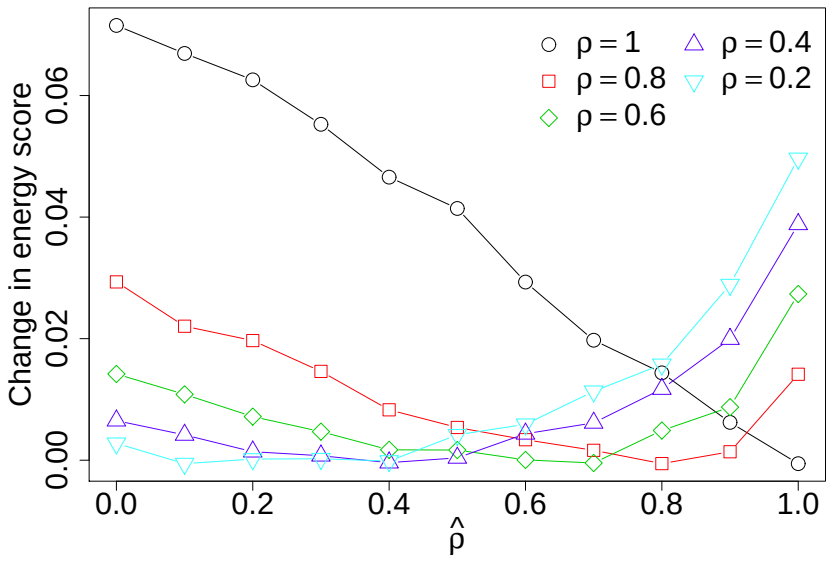


Figure 3: *Discrimination ability of the Energy score assessed with ΔEs, in terms of its sensitivity to errors in correlation for bivariate Gaussian predictive densities (hence for $n = 2$).*

upper bound.

Our result is given for the multivariate Gaussian case, for any dimension n, then discussing the specific case of $n = 2$, which may still be the most common in practice.

Let us consider that the generating process G is distributed n-variate Gaussian, $\mathbf{Y} \sim \mathcal{N}(0, \mathbf{\Sigma})$ with same variance σ^2 on all dimensions, and a correlation of 1 between any of these dimensions, i.e.,

$$\mathbf{\Sigma} = \sigma^2 \mathbf{1}_{(n \times n)}, \tag{15}$$

where $\mathbf{1}_{(n \times n)}$ is a $n \times n$ matrix of ones. This definition of the generating process implies that, at time t, a process observation $\mathbf{y}$ is such that $\mathbf{y} = y\mathbf{1}_n$, where $\mathbf{1}_n$ is a n-dimensional vector of ones, and with $y \sim \mathcal{N}(0, \sigma^2)$.

In the following, we compare the two cases where *(i)* the forecaster issues a *perfect* forecast $F = G$, and where *(ii)* the forecaster issues a so-called *naive* forecast that ignores the interdependence structure of $\mathbf{Y}$, though with appropriate mean and variance on all dimensions. In that latter case, the predictive density F is a n-variate Gaussian density with zero mean and diagonal covariance matrix, $\hat{\mathbf{\Sigma}} = \hat{\sigma}\text{diag}(\mathbf{1}_n)$. It is referred to as naive for simplicity only, since already rightly predicting mean and variance of n-variate random variables would be a nice achievement. In both cases, a closed-form formula for the Energy score is provided. They will be denoted by Es^* and Es_i. Looking at this case yield on upper bound on the discrimination ability of the Energy score for varying dependence structures, since comprising a worst case on the distance between multivariate Gaussian densities (as discussed in the above).

3.0.1 Expected Energy score for the naive forecast

For the naive forecast, one can directly work with the expression in (6). The computation of $\overline{\text{Es}_i}$ is split into that of $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X} - \mathbf{Y}||\right]\right]$ and that of $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X} - \mathbf{X}'||\right]\right]$.

After some algebra described in Appendix .1, one obtains that

$$\mathbb{E}_G\left[\mathbb{E}_F\left[||\mathbf{X} - \mathbf{Y}||\right]\right] = \frac{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)\sqrt{2}\sigma}{\Gamma\left(\frac{n}{2}\right)\sqrt{n+1}}\ {}_2F_1\left(\frac{n+1}{2}, \frac{1}{2}; \frac{n}{2}; \frac{n}{n+1}\right), \tag{16}$$

where ${}_2F_1$ is the hypergeometric function. In parallel,

$$\mathbb{E}_G\left[\mathbb{E}_F\left[||\mathbf{X}-\mathbf{X}'||\right]\right] = 2\sigma\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \tag{17}$$

The final formula for $\overline{\mathrm{Es}}_i$ therefore reads

$$\overline{\mathrm{Es}}_i = \sigma\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left\{\frac{\sqrt{2}}{\sqrt{n+1}}\,{}_2F_1\left(\frac{n+1}{2},\frac{1}{2};\frac{n}{2};\frac{n}{n+1}\right)-1\right\}. \tag{18}$$

3.0.2 Expected Energy score for the perfect forecast

In the case where $F = G$, the expression for the expected Energy score in (6) is such that

$$\overline{\mathrm{Es}}^* = \mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{Y}||\right]-\frac{1}{2}\mathbb{E}_G\left[||\mathbf{X}-\mathbf{X}'||\right]\right]. \tag{19}$$

Consequently, the calculation of the above can be split into that of $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{Y}||\right]\right]$ and that of $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{X}'||\right]\right]$.

After some algebra described in Appendix .2, one obtains that

$$\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{Y}||\right]\right] = 2\sigma\sqrt{\frac{n}{\pi}}, \tag{20}$$

while, similarly,

$$\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{X}'||\right]\right] = 2\sigma\sqrt{\frac{n}{\pi}}. \tag{21}$$

The final formula for $\bar{\mathrm{Es}}^*$ therefore reads

$$\overline{\mathrm{Es}}^* = \sigma\sqrt{\frac{n}{\pi}}. \tag{22}$$

3.1 An upper bound on the discrimination ability of the Energy score in the multivariate Gaussian case

As a consequence of the developments in the above, the upper bound $\Delta\mathrm{Es}^\dagger$ on the discrimination ability of the Energy score in the multivariate Gaussian case is obtained as

$$\Delta\mathrm{Es}^\dagger = \frac{\overline{\mathrm{Es}}_i - \overline{\mathrm{Es}}^*}{\overline{\mathrm{Es}}^*} = \sqrt{\frac{\pi}{n}}\left\{\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{\sqrt{2}}{\sqrt{n+1}}\,{}_2F_1\left(\frac{n+1}{2},\frac{1}{2};\frac{n}{2};\frac{n}{n+1}\right)-1\right)\right\}, \tag{23}$$

solely depending on the dimension n of the multivariate Gaussian process considered.

The evolution of this upper bound $\Delta\mathrm{Es}^\dagger$ is depicted in Fig. 4 as a function of the dimension of the multivariate Gaussian process of interest, from $n = 2$ to $n = 300$. This upper bound increases with n, even though it reaches an asymptote (of less than 15%) for higher dimensions. Taking the example of

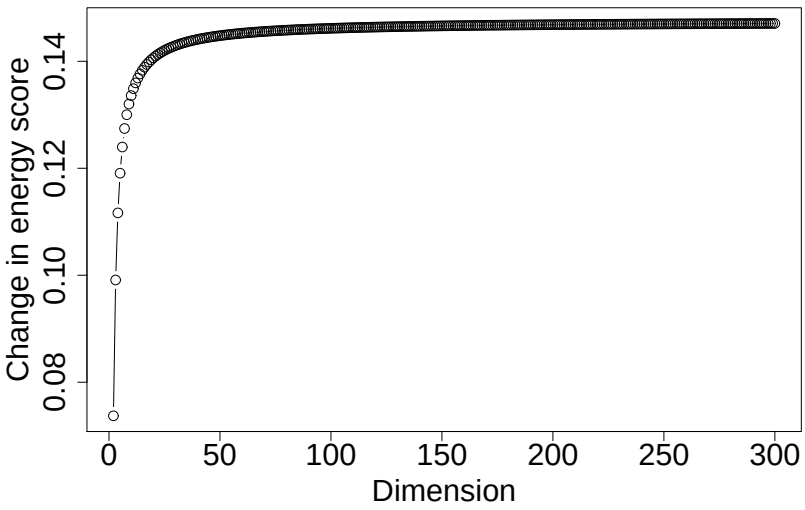


Figure 4: *Upper bound on discrimination ability (i.e., $\Delta Es^{\dagger}$) of the Energy score for a multivariate Gaussian process with well-predicted means and variances, as a function of the dimension n.*

bivariate processes, the maximum improvement in Energy score that could even be observed is of 7.4% of the Energy score value if predicting the actual distribution G of $\mathbf{Y}$. This upper bound is considerably less than if making errors in predicting the variance parameter for that multivariate Gaussian process, hence also very small compared to the case of prediction errors in the mean parameter.

Looking at Fig. 4, one could think that since the upper bound is increasing with the problem dimension, then the Energy score has a better ability to discriminate between covariance structure in higher dimensions. However, as can be seen from Fig. 3 such a theoretical upper bound is substantially higher than the differences observed under more realistic conditions for correlation values of the generating process. As the dimension of the problem grows, this upper bound may then become substantially higher than the practical discrimination ability of the Energy score.

Another important aspect to be mentioned relates to computational issues. Estimation of Energy score calls for Monte Carlo techniques, since no closed-form expression exists, even for multivariate Gaussian processes. That estimation hence becomes computationally expensive. Being more precise, the cost of sampling from a Gaussian distribution (with a covariance matrix not being restricted to any particular pattern) is cubic in the dimension. This means that estimating the Energy score is hampered by the "big n" problem.

For example in a practical application to probabilistic forecasting of wind power generation, we have considered a problem with dimension n=645 [22]. A year of hourly data was used (8760 time steps) and for each time step 10 scenarios (which is very small given the dimension of 645) were in order to evaluate the score. Given this setup, it took more than 12 hours to estimate the average Energy score when using 8 parallel cores. Such computational issues also translate to limiting the number of samples used for estimating the Energy score, therefore leading to a certain level of uncertainty in the score values obtained. This uncertainty becomes especially importance, given that a rather low sensitivity of the score to the changes in the covariance structure.

4 Discussion

The field of probabilistic forecasting is developing rapidly, and with increasing focus on multivariate processes, often of relatively low dimensions (say, $n = 2, \ldots, 5$). Considering higher dimensions will

be natural for instance for applications related to energy, meteorology and climate sciences, with focus on different variables, locations and lead times. As a consequence, it is important to further develop and analyze frameworks for probabilistic forecast verification, permitting to draw useful and practical conclusions on forecast quality.

For the case of multivariate probabilistic forecasts, the Energy score is a relevant candidate for becoming a lead score for evaluation of such forecasts. As of today, our understanding of its inherent properties for various types of processes and their varying dimensions is somewhat fairly limited. Also, the properties of related estimators, in terms of their potential bias, and sensitivity to sampling and correlation effects, are to be studied.

Our aim here was to focus on the discrimination ability of the Energy score, i.e., its ability to assign different score values to predictive densities of different quality. The most simple case of multivariate Gaussian processes and predictive densities was considered, still providing interesting insight on some of the properties of this score. Indeed, the Energy score is known to be proper, but this does not insure that it has a high discriminatory power. While it may nicely discriminate predictive densities with different mean parameters, it was discussed that differences in score values would be much less when looking at differences in their variance parameters. Also, for the case of the interdependence structure of predictive densities, an upper bound on score differences that may be observed (in expectation) was derived. Our conjecture is that, comparatively, the Energy score may hardly allow to discriminate among predictive densities with different interdependence structures. Maybe its discrimination ability could be maximized by slightly altering its definition, and using other forms of distance, better considering the structure of predictive densities. Besides, additional consideration should be given to other scoring rules defining scores for verifying multivariate probabilistic forecasts, since they may gave a different discrimination ability, while having additional computational advantages.

Acknowledgements

Acknowledgments are due to Tilmann Gneiting for various discussions on probabilistic forecast verification and properties of scoring rules.

.1 Necessary calculations to evaluate $\overline{\text{Es}}_i$

.1.1 Evaluating $\mathbb{E}_G\left[\mathbb{E}_F\left[||\mathbf{X}-\mathbf{Y}||\right]\right]$

Given a single process realization $\mathbf{y} = y\mathbf{1}_n$,

$$||\mathbf{X}-\mathbf{y}|| = \sqrt{(x_1-y)^2+(x_2-y)^2+\cdots+(x_n-y)^2}, \tag{24}$$

where a realization of $\mathbf{X}$ is $\mathbf{x} = [x_1 \ \dots \ x_n]^\top$.

In parallel, a known result is such that

$$\left.\begin{array}{rl} (x_1-y)\sim & \mathcal{N}(-y,\sigma^2) \\ (x_2-y)\sim & \mathcal{N}(-y,\sigma^2) \\ \cdots & \\ (x_n-y)\sim & \mathcal{N}(-y,\sigma^2) \\ x_1,\ x_2,\cdots,\ x_n & \text{are i.i.d.} \end{array}\right\} \Rightarrow z = \sum_{i=1}^{n}\frac{(x_i-y)^2}{\sigma^2} \sim \text{Non-central Chi-squared.}$$

Consequently, following [23], the parameters of the non-central Chi-squared distribution are given by n and λ, where

$$\lambda = \frac{1}{2}\sum_{i=1}^{n}\frac{y^2}{\sigma^2}. \tag{25}$$

Then,

$$\mathbb{E}_F\left[||\mathbf{X}-\mathbf{y}||\right] = \sigma\mathbb{E}_F\left[z^{\frac{1}{2}}\right]. \tag{26}$$

That is, in order to evaluate $\mathbb{E}_F\left[||\mathbf{X}-\mathbf{y}||\right]$ we need to know a fractional moment of order 1/2 of the non-central Chi-squared distributed variable z.

Following [23],

$$\begin{aligned}\mathbb{E}_F\left[z^{\frac{1}{2}}\right] &= \sqrt{2}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\,{}_1F_1\left(-\frac{1}{2};\frac{n}{2};-\lambda\right)\\ &= \sqrt{2}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\,{}_1F_1\left(-\frac{1}{2};\frac{n}{2};-\frac{n}{2}\frac{y^2}{\sigma^2}\right),\end{aligned} \tag{27}$$

where ${}_1F_1$ denotes the confluent hypergeometric function of the first kind. This yields

$$\mathbb{E}_F\left[||\mathbf{X}-\mathbf{y}||\right] = \sigma\sqrt{2}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\,{}_1F_1\left(-\frac{1}{2};\frac{n}{2};-\frac{n}{2}\frac{y^2}{\sigma^2}\right). \tag{28}$$

As a consequence,

$$\begin{aligned}\mathbb{E}_G\left[\mathbb{E}_F\left[||\mathbf{X}-\mathbf{Y}||\right]\right] &= \sqrt{2}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\int_{-\infty}^{\infty}{}_1F_1\left(-\frac{1}{2};\frac{n}{2};-\frac{n}{2}\frac{y^2}{\sigma^2}\right)\frac{1}{\sigma\sqrt{2\pi}}\exp\frac{-y^2}{2\sigma^2}dy\\ &= \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\sigma\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}\exp\left(-\frac{1}{2\sigma^2}y^2\right){}_1F_1\left(-\frac{1}{2};\frac{n}{2};\frac{-n}{2\sigma^2}y^2\right)\left(y^2\right)^{-\frac{1}{2}}dy^2\\ &= \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\sigma\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}\exp\left(-\frac{1}{2\sigma^2}t\right){}_1F_1\left(-\frac{1}{2};\frac{n}{2};\frac{-n}{2\sigma^2}t\right)\left(t\right)^{-\frac{1}{2}}dt\end{aligned}$$

To integrate the expression further we use (4) on page 822 of [24] stating that

$$\int_0^{\infty}\exp(-st)t^{b-1}\,{}_1F_1(a;c;kt)dt = \Gamma(b)(s-k)^{-b}F(c-a,b;c;\frac{k}{k-s}), \tag{29}$$

if $|s-k| > |k|$ and $Re(b) > 0$, $Re(s) > \max(0, Re(k))$. In the above F is the Gauss hypergeometric function.

In our case: $a = -0.5$, $c = 0.5n$, $k = \frac{-n}{2\sigma^2}$, $b = 0.5$, $s = \frac{1}{2\sigma^2}$. Then:

$|s-k| = \frac{n+1}{2\sigma^2} > \frac{n}{\sigma^2} = |k|$

$Re(b) = 0.5 > 0$

$Re(s) = \dfrac{1}{2\sigma^2} > 0 = \max(0, Re(k))$

All the conditions are fulfilled, therefore we can apply the formula given in [24] and as a result:

$$\begin{aligned}\mathbb{E}_G\left[\mathbb{E}_F\left[||\mathbf{X}-\mathbf{Y}||\right]\right] &= \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\sigma\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}\Gamma\left(\frac{1}{2}\right)\left(\frac{n+1}{2\sigma^2}\right)^{-\frac{1}{2}}\ {}_2F_1\left(\frac{n+1}{2},\frac{1}{2};\frac{n}{2};\frac{n}{n+1}\right)\\ &= \sigma\sqrt{\frac{2}{n+1}}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\ {}_2F_1\left(\frac{n+1}{2},\frac{1}{2};\frac{n}{2};\frac{n}{n+1}\right) \qquad (30)\end{aligned}$$

.1.2 Evaluating $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{X}'||\right]\right]$

As a starting point one has

$$\mathbb{E}_G\left[||X-X'||\right] = \sqrt{(x_1-x_1')^2+(x_2-x_2')^2+\cdots+(x_n-x_n')^2}, \qquad (31)$$

with $x_i,\ x_i' \sim \mathcal{N}(0,\sigma^2)$, while being mutually independent for all $i = 1, 2, \cdots, n$.

Let us introduce $z = \sum_{i=1}^{n}(x_i - x_i')^2$. Since $\dfrac{(x_i - x_i')}{2\sigma^2} \sim \mathcal{N}(0, 2\sigma^2)$, z follows a non-central Chi-squared distribution with parameters n and $\lambda = 0$.

Therefore following [23],

$$\begin{aligned}\mathbb{E}_G\left[z^{\frac{1}{2}}\right] &= \sqrt{2}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\ {}_1F_1\left(-\frac{1}{2};\frac{n}{2};0\right)\\ &= \sqrt{2}\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \qquad (32)\end{aligned}$$

Consequently,

$$\mathbb{E}_G\left[||\mathbf{X}-\mathbf{X}'||\right] = \mathbb{E}\left[z^{\frac{1}{2}}\right] = 2\sigma\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \qquad (33)$$

then also defining $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{X}'||\right]\right]$.

.2 Necessary calculations to evaluate $\overline{\mathbf{Es}}^*$

.2.1 Evaluating $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X}-\mathbf{Y}||\right]\right]$

First of all, one has

$$\mathbf{X} - y = [x_1 - y\ x_2 - y\ \cdots x_n - y]^\top. \qquad (34)$$

Since both $\mathbf{X}$ and $\mathbf{Y}$ are distributed $\mathcal{N}(0, \mathbf{\Sigma})$, with $\mathbf{\Sigma}$ as defined in (15), then $x_n = x_{n-1} = \cdots = x_1$ (which we write x) and $y_n = y_{n-1} = \cdots = y_2 = y_1$ (which we write y). Thus,

$$||\mathbf{X} - \mathbf{y}|| = \sqrt{n(x-y)^2} = \sqrt{n}|x - y|. \tag{35}$$

Subsequently, since $x \sim \mathcal{N}(0, \sigma^2)$, then given y, $(x - y) \sim \mathcal{N}(-y, \sigma^2)$. Following this, the variable $||\mathbf{X} - y||$ follows a folded Normal with parameters $-y$ and σ^2. Using the analytical expression for the mean of a folded Normal distribution (see Section .3.1), we obtain

$$\mathbb{E}_G\left[||\mathbf{X} - y||\right] = \sigma\sqrt{2n/\pi}\exp\left(\frac{-y^2}{2\sigma^2}\right) + y\left(1 - 2\Phi\left(\frac{-y}{\sigma}\right)\right). \tag{36}$$

Then given that $y \sim \mathcal{N}(0, \sigma^2)$,

$$\begin{aligned}
\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X} - \mathbf{Y}||\right]\right] &= \int_{-\infty}^{\infty}\sqrt{n}\left(\sigma\sqrt{2/\pi}\exp\left(\frac{-y^2}{2\sigma^2}\right) + y\left(1 - 2\Phi\left(\frac{-y}{\sigma}\right)\right)\right)\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-y^2}{2\sigma^2}\right)dy \\
&= \int_{-\infty}^{\infty}\frac{\sqrt{n}}{\pi}\exp\left(\frac{-y^2}{\sigma^2}\right)dy \\
&+ \sqrt{n}\sigma\int_{-\infty}^{\infty}\frac{y}{\sigma}\phi(\frac{y}{\sigma})d\frac{y}{\sigma} \\
&+ -2\sqrt{n}\sigma\int_{-\infty}^{\infty}\frac{y_1}{\sigma}\phi(\frac{y}{\sigma})\Phi(\frac{-y}{\sigma})d\frac{y}{\sigma} \\
&= \frac{\sqrt{n}\sigma}{\sqrt{\pi}} + \frac{2\sqrt{n}\sigma}{2\sqrt{\pi}} \\
&= \frac{2\sqrt{n}\sigma}{\sqrt{\pi}},
\end{aligned} \tag{37}$$

based on integrals given in Section .3.2.

.2.2 Evaluating $\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X} - \mathbf{X}'||\right]\right]$

Similarly to the above, one starts with

$$||\mathbf{X} - \mathbf{X}'|| = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2}. \tag{38}$$

Then, since $\mathbf{X},\ \mathbf{X}' \sim \mathcal{N}(0, \mathbf{\Sigma}$, with $\mathbf{\Sigma}$ as defined in (15), the above can be reformulated as

$$||\mathbf{X} - \mathbf{X}'|| = \sqrt{n(x - x')^2} = \sqrt{n}|x - x'|, \tag{39}$$

with x and x' independent draws from $\mathbf{X}$ and X' such that $\mathbf{X}, \mathbf{X}' \sim \mathcal{N}(0, \sigma^2)$. Consequently, following an argument similar to that in the previous section, it is that $|\mathbf{X} - \mathbf{X}'|$ follows a folded Normal distribution with parameters 0 and $2\sigma^2$.

By applying the formula for finding the expected value of a folded Normal density (Section .3.1), we finally obtain

$$\mathbb{E}_G\left[|\mathbf{X} - \mathbf{X}'|\right] = 2\sigma\sqrt{\frac{1}{\pi}}, \tag{40}$$

and then

$$\mathbb{E}_G\left[\mathbb{E}_G\left[||\mathbf{X} - \mathbf{X}'||\right]\right] = \frac{2\sqrt{n}\sigma}{\sqrt{\pi}}. \tag{41}$$

.3 Some results on relevant distributions and integrals

Below are given some basic definitions and results for some relevant probability distributions and integrals used in the above mathematical derivations.

.3.1 Folded Normal distribution

The folded Normal distribution is directly linked to the Normal distribution. Indeed, in the case for which X is distributed Gaussian, $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = |X|$ follows a folded Normal distribution, $Y \sim \mathcal{N}^f(\mu, \sigma^2)$. For such a distribution, the expectation of Y is given by

$$\mathbb{E}[Y] = \sigma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2\right) + \mu\left(1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right). \tag{42}$$

.3.2 Some known relevant integrals

$$\int_{-\infty}^{\infty} x\phi(x)\Phi(bx)\,dx = \int_{-\infty}^{\infty} x\phi(x)\Phi(bx)^2\,dx = \frac{b}{\sqrt{2\pi(1+b^2)}} \tag{43}$$

$$\int_{-\infty}^{\infty} e^{-ax^2}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}} \quad (a > 0) \tag{44}$$

References

[1] A. S. Tay and K. F. Wallis, "Density forecasting: a survey," *Journal of Forecasting, pages=235–254, year=2000.*

[2] A. Timmermann, "Density forecasting in economics and finance," *Journal of Forecasting*, vol. 19, no. 4, pp. 231–234, 2000.

[3] Y. Y. Kagan and D. D. Jackson, "Probabilistic forecasting of earthquakes," *Geophysical Journal International*, vol. 143, no. 2, pp. 438–453, 2000.

[4] M. Leutbecher and T. N. Palmer, "Ensemble forecasting," *Journal of Computational Physics*, vol. 227, no. 7, pp. 3515–3539, 2008.

[5] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, 2009.

[6] P. Pinson, "Wind energy: Forecasting challenges for its operational management," *Statistical Science, in press*, 2013.

[7] H. L. Cloke and F. Pappenberger, "Ensemble flood forecasting: a review," *Journal of Hydrology*, vol. 375, no. 3, pp. 613–626, 2009.

[8] T. Gneiting, "Making and evaluating point forecasts," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 746–762, 2011.

[9] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007.

[10] A. P. Dawid, "Statistical theory: The prequential approach," *Journal of the Royal Statistical Society. Series A (General)*, pp. 278–292, 1984.

[11] F. X. Diebold, A. T. Gunther, and A. S. Tay, "Evaluating density forecasts with applications to financial risk management," *International Economic Review*, vol. 39, pp. 863–883, 1984.

[12] R. Benedetti, "Scoring rules for forecast verification," *Monthly Weather Review*, vol. 138, no. 1, pp. 203–211, 2010.

[13] T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson, "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds," *Test*, vol. 17, no. 2, pp. 211–235, 2008.

[14] P. Pinson, "Adaptive calibration of (u, v)-wind ensemble forecasts," *Quarterly Journal of the Royal Meteorological Society*, vol. 138, pp. 1273–1284, 2012.

[15] D. S. Wilks, "Probabilistic canonical correlation analysis forecasts, with application to tropical Pacific sea-surface temperatures," *International Journal of Climatology*, 2013.

[16] N. Schuhen, T. L. Thorarinsdottir, and T. Gneiting, "Ensemble model output statistics for wind vectors," *Monthly Weather Review, available online*, 2012.

[17] A. Möller, A. Lenkoski, and T. L. Thorarinsdottir, "Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas," *Quarterly Journal of the Royal Meteorological Society*, 2013.

[18] P. Pinson and R. Girard, "Evaluating the quality of scenarios of short-term wind power generation," *Applied Energy*, vol. 96, pp. 12–20, 2012.

[19] Ò. Jordà and M. Marcellino, "Path forecast evaluation," *Journal of Applied Econometrics*, vol. 25, no. 4, pp. 635–662, 2010.

[20] J. Bröcker and L. A. Smith, "Scoring probabilistic forecasts: The importance of being proper," *Weather and Forecasting*, vol. 22, no. 2, pp. 382–388, 2007.

[21] A. H. Murphy, "What is a good forecast? An essay on the nature of goodness in weather forecasting," *Weather and forecasting*, vol. 8, no. 2, pp. 281–293, 1993.

[22] J. Tastu and P. Pinson, "Space-time scenarios of wind power generation produced using a Gaussian copula with parametrized precision matrix," tech. rep., Technical University of Denmark, 2013.

[23] J. R. Harvey, *Fractional moments of a quadratic form in noncentral normal random variables*. PhD thesis, North Carolina State University at Raleigh, 1965.

[24] A. Jeffrey and D. Zwillinger, *Table of integrals, series, and products*. Academic Press, 2007.