# Supplementary material to: Focusing on regions of interest in forecast evaluation

Hajo Holzmann*
Fachbereich Mathematik und Informatik
Philipps-Universität Marburg
holzmann@mathematik.uni-marburg.de

Bernhard Klar
Institut für Stochastik
Karlsruher Institut für Technologie (KIT)
bernhard.klar@kit.edu

August 7, 2017

**Abstract**

We present supplementary material to Holzmann and Klar (2017). In particular, we introduce a weighted version of the Hyvärinen score, give the proof Theorem 3 and present additional simulation results.

*Keywords.* financial time series; forecasting; logarithmic score; risk management; scoring rules; weight function

## 1 Weighted scoring rules

We start by presenting a weighted version of the Hyvärinen score from Hyvärinen (2005), which is an example of a local scoring rule. Let us recall the corresponding result from Holzmann and Klar (2017). For $p \in \mathcal{P}$, $w \in \mathcal{W}$ we let

$$p_w(x) = \frac{w(x)\, p(x)}{\int w\, p}$$

denote the renormalized density of $p$ w.r.t. $w$. For formulating the next result, let $\tilde{\mathcal{P}}$ be another class of densities such that $p_w \in \tilde{\mathcal{P}}$ for every $w \in \mathcal{W}$, $p \in \mathcal{P}$.

**Theorem 1.** *Let $\widetilde{S} : \widetilde{\mathcal{P}} \times \mathcal{X} \to \overline{\mathbb{R}}$ be a proper scoring rule. Then*

$$S : \mathcal{P} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}, \quad S(p, x; w) = w(x)\, \widetilde{S}(p_w, x)$$

*is a localizing proper weighted scoring rule. Further, if $\widetilde{S}$ is strictly proper, then $S$ is proportionally locally proper.*

*Corresponding author. Prof. Dr. Hajo Holzmann, Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Hans-Meerweinstr., 35043 Marburg, Germany

In the following example, Theorem 1 is applied to the Hyvärinen score from Hyvärinen (2005), resulting in a proportionally proper localizing scoring rule which is also a local scoring rule in the sense of Ehm and Gneiting (2012). It is insensitive to the normalizing constant, but in contrast to the conditional likelihood score, it can be evaluated without knowing the normalizing constant of the density forecasts.

**Example** (*local scoring rules and weighting*)**.** The localizing proper weighted scoring rules from Holzmann and Klar (2017) should not be confused with *proper local scoring rules*, as investigated by Ehm and Gneiting (2012) and Parry et al. (2012). As in these papers, consider first the real-valued situation. The dominating measure is the Lebesgue measure, and a scoring rule is local of order $k$ if $S(p, x)$ only depends on $p$ through the first $k$ derivatives $p(x), p'(x), \ldots, p^{(k)}(x)$ of $p$ at $x$. Parry et al. (2012) show existence of proper local scoring rules for any given even order, while proper scoring rules of odd order do not exist. Ehm and Gneiting (2012) characterize proper local scoring rules of order two.
In order to apply Theorem 1 to higher-order local scoring rules, we require the weight function to be sufficiently smooth as well. This excludes indicator functions, but allows for smooth approximations of indicators. Now, proper, higher-order local scoring rules can be computed without normalizing the density $p$, and thus we expect that the factor $\int pw$ in $\widehat{S}(p, x; w)$ in Theorem 1 cancels out, leaving us again with a proper local scoring rule of the same order.
Let us illustrate this for the local scoring rule introduced by Hyvärinen (2005),

$$S(p, x) = 2\,\frac{p''(x)}{p(x)} - \Big(\frac{p'(x)}{p(x)}\Big)^2.$$

Under conditions on the class of densities stated in Ehm and Gneiting (2012) or Hyvärinen (2005), it is strictly proper. When applying Theorem 1, after canceling terms which only depend on $w$ but not on the forecast density $p$, and dividing by two, we get

$$\widehat{S}(p, x; w) = 2\,\frac{p''(x)}{p(x)}\,w(x) - \Big(\frac{p'(x)}{p(x)}\Big)^2 w(x) + 2\,\frac{p'(x)}{p(x)}\,w'(x), \tag{1}$$

for which

$$\widehat{S}(p, q; w) - \widehat{S}(q, q; w) = \int \Big[\frac{p'}{p} - \frac{q'}{q}\Big]^2 q\,w.$$

The rule $\widehat{S}$ is also a local scoring rule of order two, and belongs to the class characterized by Ehm and Gneiting (2012). Indeed, in their Theorem 3.2, if we put $c = 0$ and $K_0(x, y_1) = -w(x)\,y_1^2$, then

$$s(x, y_0, y_1, y_2) = w(x)\,(y_1^2 + 2y_2) + 2w'(x)y_1.$$

Observing

$$y_1 = \frac{p'}{p}, \qquad y_2 = \frac{p''}{p} - \Big(\frac{p'}{p}\Big)^2$$

gives the weighted Hyvärinen score (1). Note that, as a locally proper weighted scoring rule, $\widehat{S}$ is also *preference preserving* in the sense introduced by Pelenis (2014) since it depends linearly on the weight function $w$.
We can also obtain a multivariate version by applying Theorem 1 to the multivariate score of Hyvärinen (2005). For $(\mathcal{X}, \mathcal{F}) = (\mathbb{R}^d, \mathcal{B}^d)$, using the notation

$$\partial_i\, p(\mathbf{x}) := \frac{\partial}{\partial\, x_i} p(\mathbf{x}), \qquad \partial_i^2\, p(\mathbf{x}) := \Big(\frac{\partial}{\partial\, x_i}\Big)^2 p(\mathbf{x}), \quad i = 1, \ldots, d,$$

we obtain

$$\widehat{S}(p, \mathbf{x}; w) = \sum_{i=1}^{d} \Big[ 2 \, \frac{\partial_i^2 \, p(\mathbf{x})}{p(\mathbf{x})} \, w(\mathbf{x}) - \Big( \frac{\partial_i \, p(\mathbf{x})}{p(\mathbf{x})} \Big)^2 \, w(\mathbf{x}) + 2 \, \frac{\partial_i^2 \, p(\mathbf{x})}{p(\mathbf{x})} \, \partial_i w(\mathbf{x}) \Big]$$

after cancelling terms not depending on $p$ and dividing by two, for which we have that

$$\widehat{S}(p, q; w) - \widehat{S}(q, q; w) = \int_{\mathbb{R}^d} \Big[ \sum_{i=1}^{d} \Big( \frac{\partial_i \, p(\mathbf{x})}{p(\mathbf{x})} - \frac{\partial_i \, q(\mathbf{x})}{q(\mathbf{x})} \Big)^2 \Big] \, q(\mathbf{x}) \, w(\mathbf{x}) \, d\mathbf{x}.$$

The fact that $\widehat{S}(p, \mathbf{x}; w)$ does not depend on the normalization constant $\int pw$ may be useful in high-dimensional applications, where computation of $\int pw$ can be difficult. However, as a consequence, $\widehat{S}(p, \mathbf{x}; w)$ is only proportionally locally proper and not strictly locally proper. ◇

Next, we state and prove Theorem 3 in Holzmann and Klar (2017).

**Theorem 3.** *(i) Let $\widetilde{S} : \widetilde{\mathcal{M}} \times \mathcal{X} \to \overline{\mathbb{R}}$ be a proper scoring rule. Then*

$$S : \mathcal{M} \times \mathcal{X} \times \mathcal{W} \to \overline{\mathbb{R}}, \quad S(P, x; w) = w(x) \, \widetilde{S}(P_w, x)$$

*is a localizing proper weighted scoring rule. Further, if $\widetilde{S}$ is strictly proper, then $S$ is proportionally locally proper.*

*(ii) Let $\mathbf{s}(\alpha, z)$ be a strictly proper scoring rule for the success probability $\alpha \in (0, 1)$ of a binary outcome variable $z \in \{0, 1\}$. Then*

$$S_{\mathbf{s}}(P, x; w) = w(x) \, \mathbf{s}\Big( \int w \, \mathrm{d}P, 1 \Big) + \big( 1 - w(x) \big) \, \mathbf{s}\Big( \int w \, \mathrm{d}P, 0 \Big). \tag{2}$$

*is a localizing proper weighted scoring rule for the probability forecast $P$.*

*(iii) If $S_{\mathbf{s}}(P, x; w)$ is as in (2) and if $S(P, x; w)$ is a proportionally locally proper weighted scoring rule, then*

$$\widehat{S}(P, x; w) = S_{\mathbf{s}}(P, x; w) + S(P, x; w)$$

*is strictly locally proper.*

*Proof of Theorem 3.* (i) By Definition, $S(P, x; w)$ depends on $P$ only through $P_w$, and hence on $P$ only on $\{w > 0\}$, thus, $S(P, x; w)$ is localizing. Further, we have

$$\begin{aligned} S(P, Q; w) &= \int \widetilde{S}(P_w, x) \, \mathrm{d}Q_w(x) \int w \, \mathrm{d}Q \\ &= \widetilde{S}(P_w, Q_w) \int w \, \mathrm{d}Q. \end{aligned}$$

Since $\widetilde{S}$ is proper, $S(P, Q; w)$ is minimal in $P$ for given $Q$ if $P_w = Q_w$, which is implied by $P = Q$. Hence $S(P, x; w)$ is proper. Further, $S(P, Q; w) = S(Q, Q; w)$ implies that $\widetilde{S}(P_w, Q_w) = \widetilde{S}(Q_w, Q_w)$. Hence, if $\widetilde{S}$ is strictly proper, this implies that $P_w = Q_w$. We argue that this holds if and only if

$$\forall \, F \in \mathcal{F} : \; P\big( \{w > 0\} \cap F \big) = c \, Q\big( \{w > 0\} \cap F \big), \tag{3}$$

where necessarily

$$c = \frac{\int w \, \mathrm{d}P}{\int w \, \mathrm{d}Q}. \tag{4}$$

Indeed, we have that

$$\forall\, F \in \mathcal{F}:\ P\big(\{w>0\}\cap F\big) = \int_{F\cap\{w>0\}} \frac{\int w\,\mathrm{d}P}{w(x)}\,\mathrm{d}P_w,$$
$$\forall\, F \in \mathcal{F}:\ Q\big(\{w>0\}\cap F\big) = \int_{F\cap\{w>0\}} \frac{\int w\,\mathrm{d}Q}{w(x)}\,\mathrm{d}Q_w.$$

Hence if $P_w = Q_w$ this certainly implies (3), where $c$ is give in (4). On the other hand, if (3) holds for some $c>0$, then by approximation with step functions it follows that for any measurable function $g\geq 0$ with $\{g>0\}\subset\{w>0\}$, we have that $\int g\,\mathrm{d}P = c\int g\,\mathrm{d}Q$. Choosing $g=w$ gives (4) for $c$. Given $F\in\mathcal{F}$, to show $P_w(F)=Q_w(F)$ we then choose $g(x)=w(x)\,1\{x\in F\}$.

(ii) The rule $S_{\mathbf{s}}$ is localizing w.r.t. $P$ since it depends only on $\int w\,\mathrm{d}P$. Further, it is proper since

$$S_{\mathbf{s}}(P,Q;w) - S_{\mathbf{s}}(Q,Q;w) = \mathbf{s}\Big(\int w\,\mathrm{d}P, \int w\,\mathrm{d}Q\Big) - \mathbf{s}\Big(\int w\,\mathrm{d}Q, \int w\mathrm{d}Q\Big) \geq 0, \tag{5}$$

where we used the notation

$$\mathbf{s}(\alpha,\beta) = \beta\mathbf{s}(\alpha,1) + (1-\beta)\mathbf{s}\big(\alpha,0\big).$$

(iii) As a sum of two locally proper scoring rules the rule $\widehat{S}$ is also a locally proper scoring rule. Further, if

$$\widehat{S}(Q,Q;w) = \widehat{S}(P,Q;w),$$

then necessarily $S(Q,Q;w)=S(P,Q;w)$ and $S_{\mathbf{s}}(Q,Q;w)=S_{\mathbf{s}}(P,Q;w)$ since both rules $S(\cdot,\cdot;w)$ and $S_{\mathbf{s}}(\cdot,\cdot;w)$ are proper. By assumption on $S(\cdot,\cdot;w)$, $S(Q,Q;w)=S(P,Q;w)$ implies (3), and from the above discussion we necessarily have (4). From $S_{\mathbf{s}}(Q,Q;w)=S_{\mathbf{s}}(P,Q;w)$, (5) and the fact that $\mathbf{s}$ is strictly proper we get that $\int w\mathrm{d}P = \int w\mathrm{d}Q$ and hence $c=1$, so that $\widehat{S}$ is strictly locally proper. □

Theorem 3 allows to obtain weighted versions of general, possibly multivariate energy scores, which are defined in analogy to display (11) in Holzmann and Klar (2017), but for which a representation in terms of Brier scores does not exist.

Assume that $\mathcal{M}$ is a family of distributions on $\mathbb{R}^n$, and given $\beta\in(0,2]$ assume that $\mathrm{E}_P\,\|X\|^\beta<\infty$ for all $P\in\mathcal{M}$, where $\|\cdot\|$ is the Euclidean norm, and $X$ is the random vector distributed according to $P$. The *energy score* is defined by

$$\mathrm{ES}(P,x) = \mathrm{E}_P\big\|x-X\big\|^\beta - \frac{1}{2}\,\mathrm{E}_P\big\|X'-X\big\|^\beta,$$

where $X$ and $X'$ are independent copies distributed according to $P$, and $\|\cdot\|$ is the Euclidean norm. If we take $w(x)=1\{\|x\|>r\}$ and use the Brier score to complement for a strictly locally proper rule, we obtain

$$\begin{aligned}\mathrm{ESws}(P,x;r) &= 1\{x\leq r\}\,\big(\mathbb{P}_P(\|X\|>r)\big)^2 + 1\{x>r\}\,\big(1-\mathbb{P}_P(\|X\|>r)\big)^2 \\ &\quad + 1\{x>r\}\left(\frac{\mathrm{E}_P\big(\big\|x-X\big\|^\beta 1\{\|X\|>r\}\big)}{\mathbb{P}_P(\|X\|>r)} - \frac{\mathrm{E}_P\big(\big\|X'-X\big\|^\beta 1\{\min(\|X\|,\|X'\|)>r\}\big)}{2\,(\mathbb{P}_P(\|X\|>r))^2}\right).\end{aligned}$$

Note that the weighted CRPS of Gneiting and Ranjan (2011) cannot be easily generalized in this fashion.

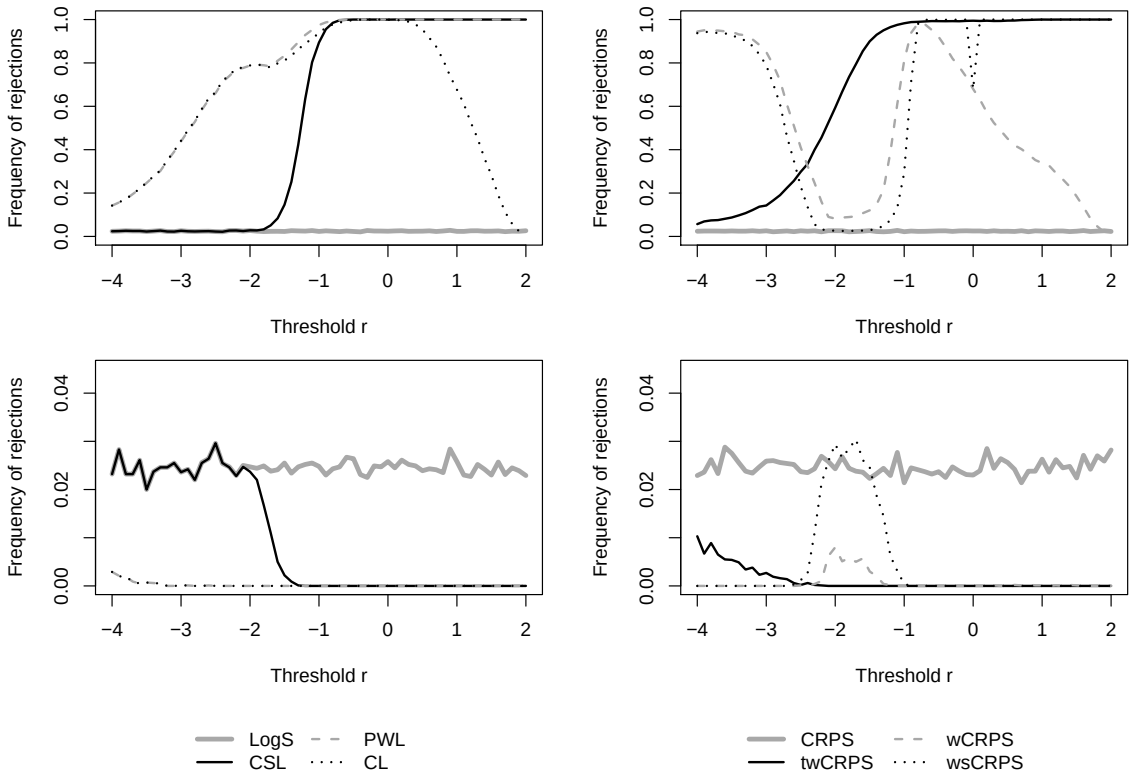


Figure 1: Scenario A. The null hypothesis of equal predictive performance of $F_{hlt}$ and $F_{hrt}$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Wilcoxon signed-rank test for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of $F_{hlt}$ (in favor of $F_{hrt}$).

# 2 Additional simulation results

In this section, we show additional results, mainly using the nonparametric Wilcoxon signed-rank test, in Scenarios A-D of the main text. As in the main text, all results in this section are based on 10 000 replications.

**Scenario A:** Forecast 1: $F_{hlt}$ vs. Forecast 2: $F_{hrt}$.
Figure 1 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Wilcoxon signed-rank test as a function of the threshold value $r$ in the weight function. The upper (lower) panels show rejections in favor of $F_{hlt}$ (in favor of $F_{hrt}$).
For $r = -\infty$, both forecasts have the same distance from the (true) standard normal distribution, and neither of them should be rejected in favor of the other. However, for $r > 0$, Forecast 1 coincides with $\Phi$, and Forecast 2 should be rejected.

Qualitatively, the results are similar as for the Diebold-Mariano test used in the main text. Main differences are: rejection rates in favor of $F_{hlt}$ are much higher, reaching 100% for $r > -1$ for the

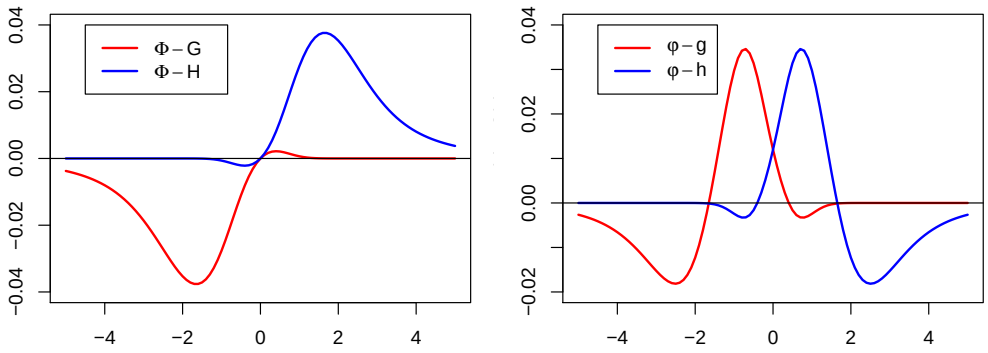


Figure 2: Scenario B. Differences of cdf's $\Phi - G$ and $\Phi - H$ (left) and differences of pdf's $\varphi - g$ and $\varphi - h$ (right).

weighted scoring rules (but again, rejection rates for CL and wCRPS decrease to zero for large positive values of $r$). Rejection rate of twCRPS increases faster than the rates of wCRPS and wsCRPS.
It is remarkable that the rejection rate of CSL in favor of $F_{hlt}$ increases faster than that of twCRPS using the Diebold-Mariano test, but vice versa in case of the Wilcoxon signed-rank test.

**Scenario B:** Forecast 1: $G$ vs. Forecast 2: $H$.
In Scenario B, both forecasts are different from the true standard normal distribution on each observation window $[r, \infty)$. For $r = -\infty$, both forecasts have the same overall distance from the standard normal distribution, and neither of them should be rejected in favor of the other. However, if one is only interested in the region $[r, \infty)$ for larger positive values of $r$, forecast $G$ is close to $\Phi$; hence, $H$ should be rejected. Figure 2 shows the differences of the cdf's $\Phi - G$ and $\Phi - H$ (left panel) and the differences of the pertaining densities $\varphi - g$ and $\varphi - h$ (right panel) to give a visual impression of the distance from $G$ and $H$ to $\Phi$.
Figure 3 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Diebold-Mariano tests as a function of the threshold value $r$ in the weight function.
Qualitatively, the results of all simulations for this scenario parallel the findings for Scenario A. Rejection frequencies in favor of $G$ and in favor of $H$ for the two non-weighted scoring rules are around 0.025. For negative values of $r$, CL and PWL behave similar, showing a faster increase of the rejection frequencies in favor of $G$ compared to CSL. However, CL decreases to zero for large positive values of $r$. Hence, PWL and CSL are clearly preferable to CL.
Concerning the CRPS based weighted scoring rules, wCRPS and wsCRPS behave quite similarly for negative and moderately positive values of $r$. Their rejection frequencies in favor of $G$ have a first modal value around $r = -2.5$, decrease until -1.5, and increase again. However, wCRPS finally decreases for large positive values of $r$.
Figure 4 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Wilcoxon signed-rank test as a function of the threshold value $r$ in the weight function. The upper (lower) panels show rejections in favor of $G$ (in favor of $H$).

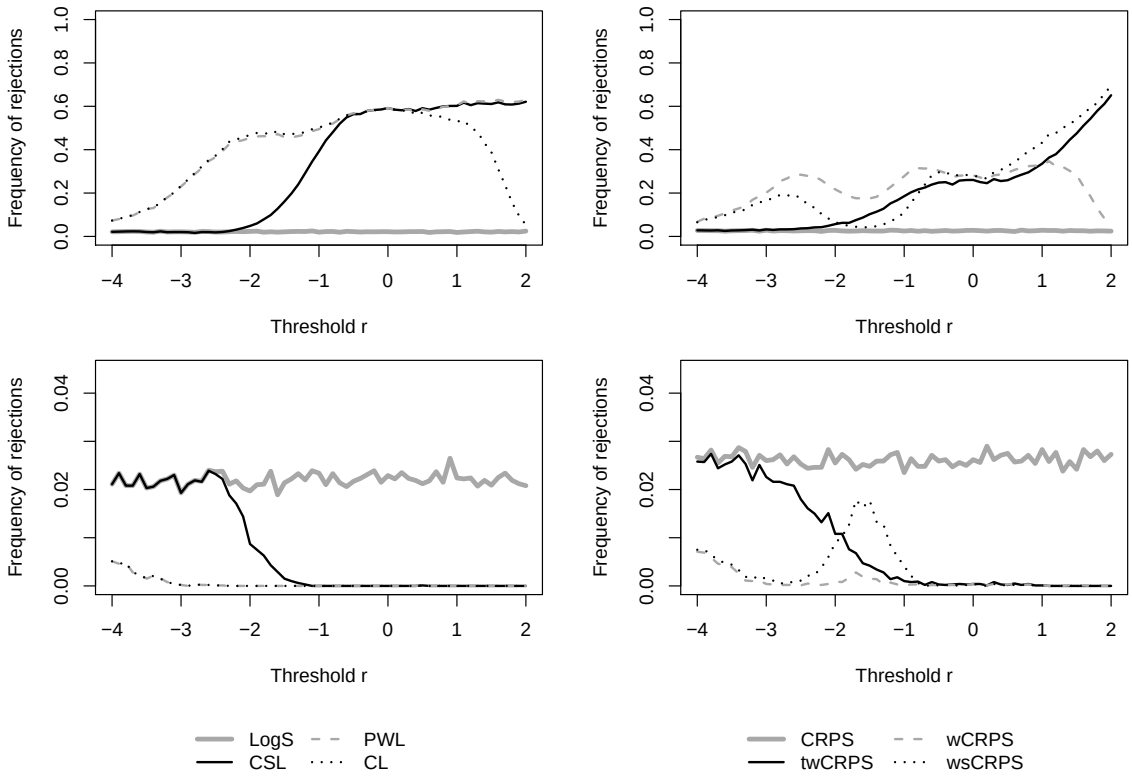


Figure 3: Scenario B. The null hypothesis of equal predictive performance of $G$ and $H$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold-Mariano tests for likelihood based (left) and CRPS based (right) scoring rules for sample size $n = 100$. The upper (lower) panels show rejections in favor of $G$ (in favor of $H$).

Again, the results are qualitatively similar as for the Diebold-Mariano test, but with much higher rejection rates in favor of $F_{hlt}$ rapidly reaching 100% for the weighted scoring rules. Rejection rate of twCRPS increases faster than the rates of wCRPS and wsCRPS.

Figures 5 and 6 are arranged in the same manner as Figures 3 and 4, respectively, but instead of the fixed sample size $n = 100$, we use sample sizes varying with the threshold value $r$ in such a way that under the standard normal distribution the expected number, $c = 10$, of observations in the relevant region $[r, \infty)$ remains constant.
Qualitatively, the panels of Figures 5 and 6 are similar to the corresponding panels for fixed sample size, but rejection frequencies in favor of $G$ are higher for large positive values of the threshold.

**Scenario C:** Forecast 1: $\Phi$ vs. Forecast 2: $F_{hlt}$.

Figure 7 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Wilcoxon signed-rank test as a function of the threshold value $r$ in the weight function. The upper (lower) panels show rejections in favor of $\Phi$ (in favor of $F_{hlt}$).

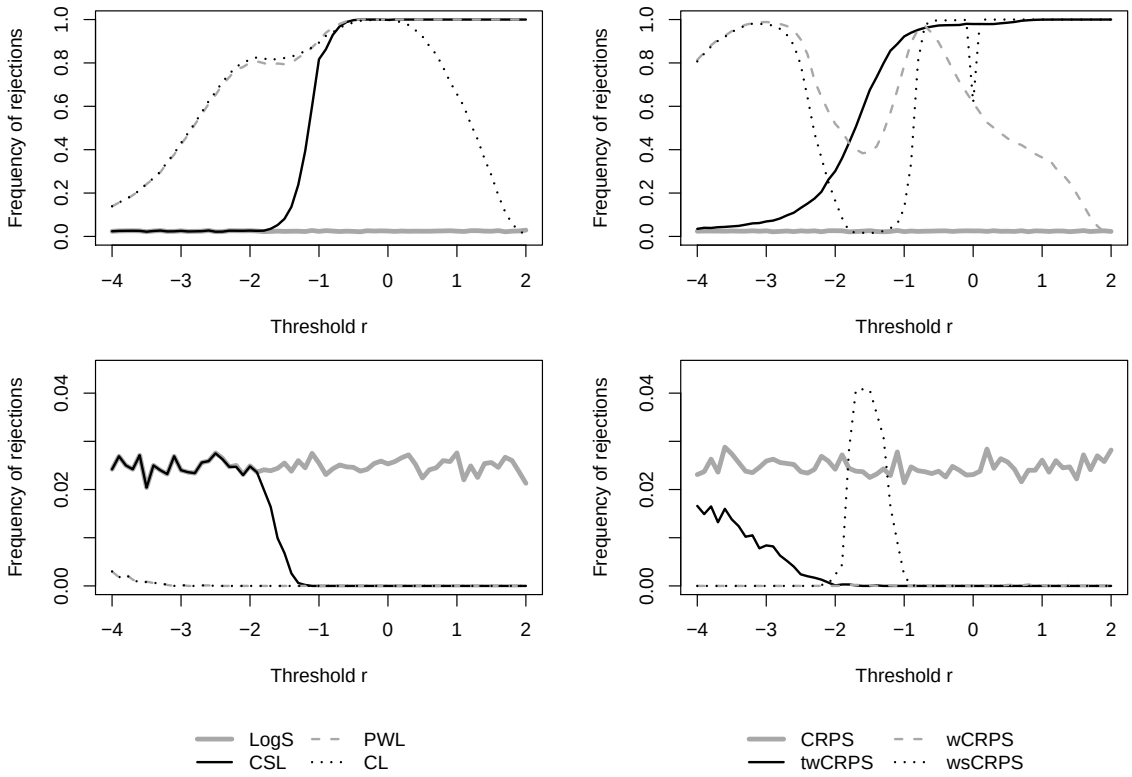


Figure 4: Scenario B. The null hypothesis of equal predictive performance of $G$ and $H$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Wilcoxon signed-rank test for likelihood based (left) and CRPS based (right) scoring rules for sample size $n = 100$. The upper (lower) panels show rejections in favor of $G$ (in favor of $H$).

For $r < 0$, rejections in favor of the standard normal distribution represent true power, but if one is interested in the region $[r, \infty)$ for positive $r$, both forecasts are identical, and neither of them should be rejected.
Looking first at the non-weighted scoring rules, we see that both rules reject $F_{hlt}$ in each case.
For the weighted scoring rules, the results are again qualitatively similar as for the Diebold-Mariano test. Again, rejection frequencies in favor of $F_{hlt}$ have a peculiar peak to the left of zero for all likelihood based weighted scoring rules, but also for twCRPS.

**Scenario D:** Forecast 1: $\Phi$ vs. Forecast 2: $G$.

Figure 8 shows the proportion of rejections of the null hypothesis of equal predictive performance in two-sided Wilcoxon signed-rank test as a function of the threshold value $r$ in the weight function. The upper (lower) panels show rejections in favor of $\Phi$ (in favor of $G$).
Formally, rejections in favor of the standard normal distribution represent true power, but if one is interested in the region $[r, \infty)$ for positive $r$, both forecasts are quite similar, and neither of them should be rejected.

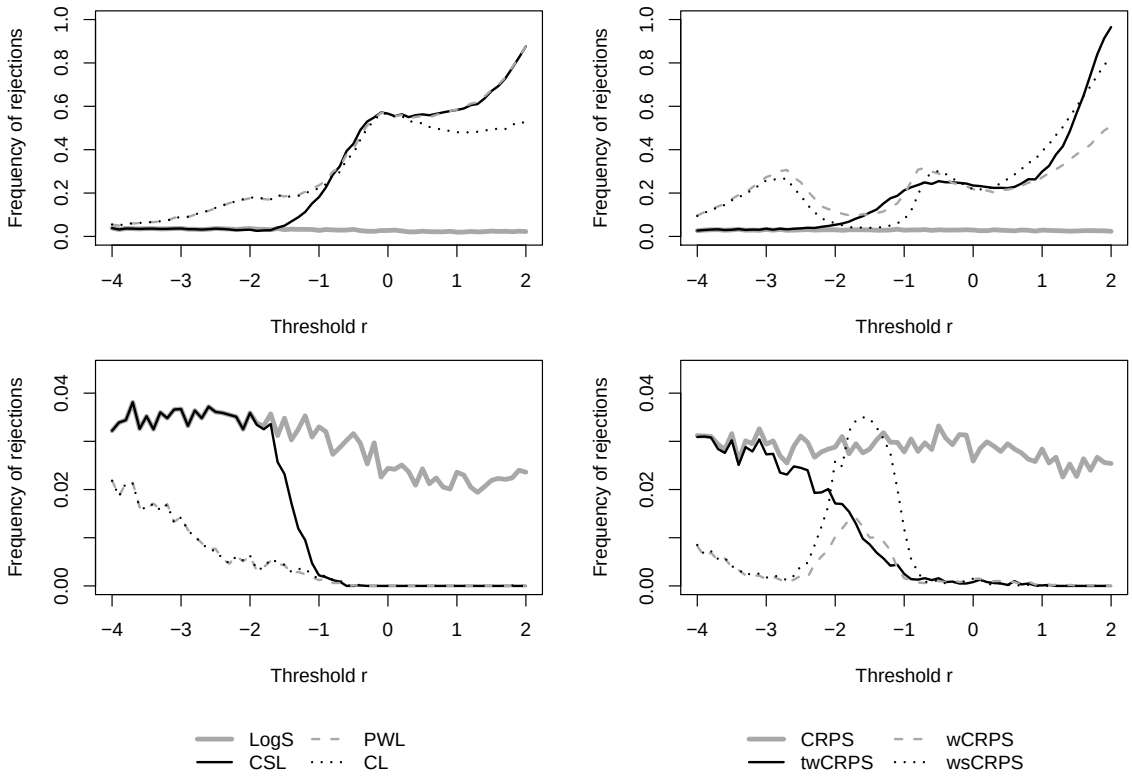


Figure 5: Scenario B. The null hypothesis of equal predictive performance of $G$ and $H$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold-Mariano test for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of $G$ (in favor of $H$). The sample size $n$ varies in such a way that the expected number of observations in $[r, \infty)$ is 10 under the standard normal distribution.

Qualitatively, most results for this scenario parallel the findings for Scenario C. Both non-weighted scoring rules reject $F_{hlt}$ in nearly each case. Looking at the weighted scoring rules, we see again a high peak in the rejection frequencies in favor of $F_{hlt}$ to the left of zero for CSL and PWL, but only to a much lesser extent for CL and twCRPS.

# References

Ehm, W. and Gneiting, T. (2012). *Local proper scoring rules of order two.* Annals of Statistics 40, 609-637.

Gneiting, T. and Ranjan, R. (2011). *Comparing density forecasts using threshold- and quantile-weighted scoring rules.* Journal of Business and Economic Statistics 29, 411-422.

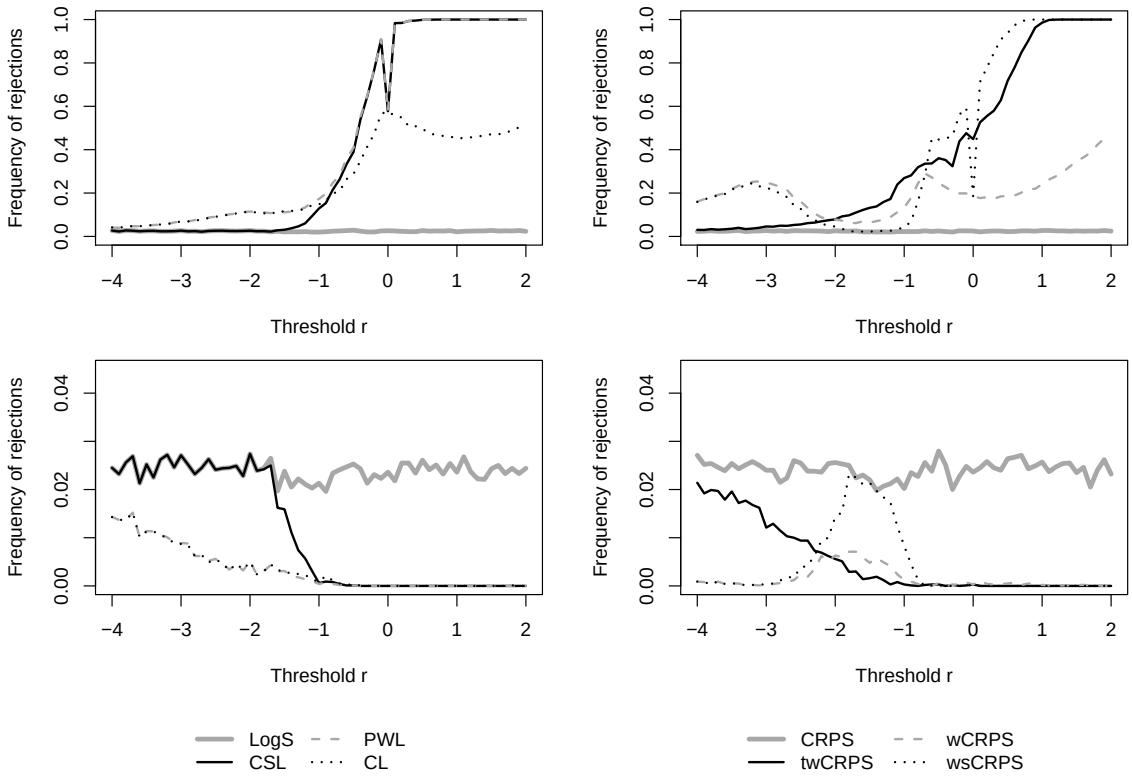


Figure 6: Scenario B. The null hypothesis of equal predictive performance of $G$ and $H$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Wilcoxon signed-rank test for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of $G$ (in favor of $H$). The sample size $n$ varies in such a way that the expected number of observations in $[r, \infty)$ is 10 under the standard normal distribution.

Holzmann, H. and Klar, B. (2017). *Focusing on regions of interest in forecast evaluation.* submitted to *Annals of Applied Statistics*

Hyvärinen, A. (2005). *Estimation of non-normalized statistical models by score matching.* Journal of Machine Learning Research 6, 695-709.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). *Proper local scoring rules.* Annals of Statistics 40, 561-592.

Pelenis, J. (2014). *Weighted scoring rules for comparison of density forecasts on subsets of interest.* Preprint. Retrieved from `https://sites.google.com/site/jpelenis/` in July, 2017.

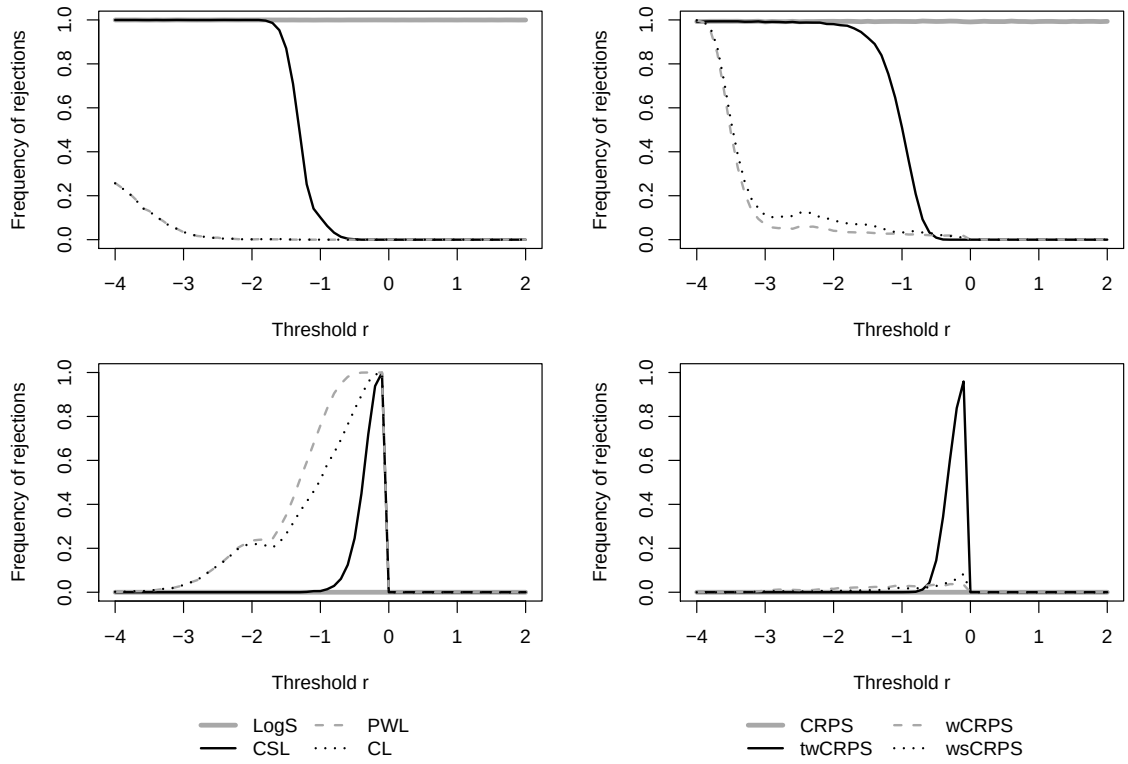


Figure 7: Scenario C. The null hypothesis of equal predictive performance of $\Phi$ and $F_{hlt}$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Wilcoxon signed-rank test for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of $\Phi$ (in favor of $F_{hlt}$).

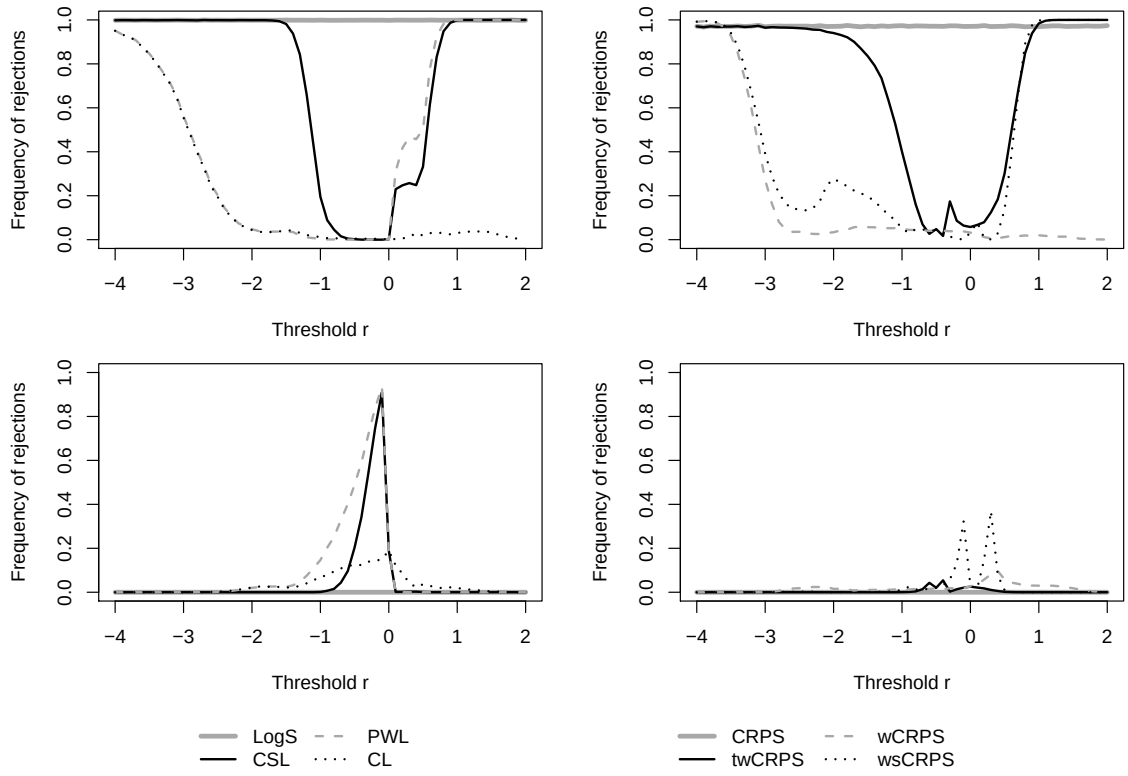


Figure 8: Scenario D. The null hypothesis of equal predictive performance of $\Phi$ and $F_{hlt}$ is tested under a standard normal population. The panels show the frequency of rejections in two-sided Wilcoxon signed-rank test for likelihood based (left) and CRPS based (right) scoring rules. The upper (lower) panels show rejections in favor of $\Phi$ (in favor of $F_{hlt}$).