Online Appendix for "Information Aggregation and P-hacking"

by Oleg Rytchkov and Xun Zhong

Alternative simulation specifications

In this section of the Online Appendix, we consider several modifications of the baseline time series setting from Section 2.2.1 and demonstrate that our qualitative conclusions are robust to them. We report the results only for the case with $T = 276$ observations in the whole sample and $T_{OS} = 140$ observations in the quasi-out-of-sample period.

Alternative number of aggregated predictors. In the first set of robustness tests we explore the sensitivity of our results to the number of aggregated predictors by repeating the analysis from Section 2.2.1 for $A = 10$, $A = 20$, and $A = 30$. Overall, the results reported in Table OA.1 are comparable to those in the middle part of Panel B of Table 1. In particular, 3PRF produces positive out-of-sample performance measures for all considered sets of spurious predictors. In contrast, all performance measures of the 3PRFm forecasts are negative. The comparison of the cases with $A = 10$, $A = 20$, and $A = 30$ delivers several new insights. First, the increase in the number of spurious predictors substantially increases the average out-of-sample R^2 and Clark and West (2007) t-statistics of the 3PRF forecasts. This result is not surprising because the amount of information about in-sample realizations of returns increases with A, and 3PRF efficiently aggregates it. Second, the impact of the increase in the number of signals on the 3PRFm forecast is drastically different. Both R^2_{OS} and Clark and West (2007) t-statistics become more negative, and the magnitude of the effect is large: for example, the absolute value of mean R^2_{OS} increases almost three times as A changes from 10 to 30. Thus, 3PRFm penalizes a larger number of spurious predictors more severely and sets a higher bar for potentially informative signals. This result is consistent with the intuition presented in Section 2.1.3. Indeed, a larger A implies a better diversification of individual signals in X^W_t. Therefore,

in such cases X_t^W is closer to X_t^E, and the difference $X_t^W - X_t^E$ in equation (2) is less informative about returns than in the cases with small A.

Alternative critical value. The inability of 3PRFm to produce a powerful forecast can be possibly explained by relative weakness of individual signals and low efficiency of the aggregation procedure. To demonstrate that this is not the case, we repeat the simulations from Section 2.2.1 but generate spurious signals postulating that the signal predicts returns in sample only if the absolute value of its t-statistic in the predictive regression exceeds 3. This critical value for individual signals is promoted in the cross-sectional analysis of stock returns by Harvey et al. (2016) as a safeguard against spurious signals.

The comparison of the left panel of Table OA.2 and the middle part of Panel B of Table 1 reveals that the individual signals with the t-statistics of 3 are indeed stronger than their counterparts in the baseline case: their 3PRF combinations are better predictors of returns with R^2_{OS} that are almost twice as large as R^2_{OS} in the baseline case. Nevertheless, the 3PRFm forecast still produces negative R^2_{OS} as well as zero probability to reject the null of no predictability. Moreover, the 3PRFm performance measures are even slightly more negative than their analogs in Table 1, and this fact can be explained by a lower dispersion in the quality of spurious signals when a higher critical value is used for their selection. Thus, as explained in Section 2.1.3, 3PRFm largely ignores signals with similar predictive abilities irrespective of how strong they are.

Correlated spurious signals. In the main analysis, we assume that spurious signals are uncorrelated. However, this assumption often does not hold, and its violation may affect our conclusions. Therefore, we repeat the analysis from Section 2.2.1 using correlated spurious signals that are constructed as follows. First, we generate a common factor in signals f_t as a time series with T observations that are random draws from the standard normal distribution. Second, we generate a vector of signal loadings δ^a, $a = 1, \ldots, A$, so that each element of the vector is a random draw from the uniform distribution $U[-0.95, 0.95]$. Third, we construct signals as $X_t^a = \delta^a f_t + u_t^a$, where u_t^a are randomly drawn from the normal distribution $\mathcal{N}(0, 1 - (\delta^a)^2)$. Note that by construction all signals have unit variance and $\text{Corr}(X_t^a, X_t^b) = \delta^a \delta^b$. Because the distribution

of δ^a, $a = 1, \ldots, A$, is symmetric around zero, the signals can be either positively or negatively correlated. Moreover, the average correlation between them is $\mathrm{E}[\mathrm{Corr}(X_t^a, X_t^b)] = \mathrm{E}[\delta^a \delta^b] = 0$, and the standard deviation of the correlations is $\sigma[\mathrm{Corr}(X_t^a, X_t^b)] = \sigma[\delta^a \delta^b] = \sigma(\delta^a)^2 = 0.95^2/3 \approx 0.3$. The moments are close to their counterparts for empirical predictors of stock returns from Section 2.4 (0.15 and 0.3, respectively).

For the sets of correlated spurious signals, the out-of-sample statistics of the 3PRF and 3PRFm forecasts are reported in the middle panel of Table OA.2. Compared to the case with uncorrelated signals, 3PRF produces lower R^2_{OS}. Indeed, all signals have a strong common factor f_t, which is unrelated to future returns, and the factor makes it harder for the combined forecasts to produce high R^2_{OS}. Most importantly, the 3PRFm forecast still has substantially negative R^2_{OS}, and the probability that the Clark and West (2007) t-statistic exceeds 1.645 is zero. Therefore, the robustness of 3PRFm to p-hacking is largely unaffected by correlations between signals.

Real returns, spurious signals. In the main analysis, we simulate returns as random draws from the normal distribution with the prespecified mean and variance. However, actual returns have a more complicated distribution. To demonstrate the robustness of our results to the distributional assumption, we repeat the simulations from Section 2.2.1 using actual realizations of returns that are chosen to be continuously compounded excess returns on the S&P 500 index, including dividends.

The results reported in the right panel of Table OA.2 show that all out-of-sample statistics are close to their counterparts in the main analysis. In particular, 3PRF produces positive R^2_{OS} and the Clark and West (2007) test always rejects the null of no predictability, but 3PRFm again fails to construct a reliable predictor of returns from spurious signals. Thus, our conclusions are insensitive to the assumptions about statistical properties of returns and hold for actual returns.

Predictability of aggregate stock returns by 13 predictors: additional tests

In this section of the Online Appendix, we report the results on the predictability of aggregate stock returns by 13 predictors in various out-of-sample periods. As documented in the literature, the strength of return predictability varies over time (e.g., Paye and Timmermann 2006; Rapach and Wohar 2006). Therefore, the inability of the 3PRFm forecast to produce positive R^2_{OS} in the out-of-sample period from January 1981 to December 2015 may not generalize to other periods. To demonstrate the robustness of the results reported in Table 4, we repeat the analysis for six alternative out-of-sample periods: three of them are 1965–2005, 1976–2005, and 2000–2005 (exactly as in Rapach et al. (2010)) and three others are extensions of those periods up to 2015. Because some of the periods are relatively short, we conduct the analysis only for quarterly and monthly returns.

Table OA.3 shows that the EW, MSFE, and DMSFE forecasts are particularly strong and statistically significant in the periods 1965–2005 and 2000–2005, and this result is consistent with Rapach et al. (2010). Moreover, it holds for both quarterly and monthly returns. The inclusion of the last 10 years in the sample typically decreases R^2_{OS}, although the qualitative conclusions remain the same. As in Table 4, the OLS forecasts consistently produce negative R^2_{OS}. The predictive ability of the PCR and 3PRF forecasts substantially varies across the samples: in the periods 1976–2005 and 1976–2015 their R^2_{OS} are negative, but in all other periods they are positive with the p-values of the Clark and West (2007) t-statistic smaller than 0.1. This observation is largely consistent across return frequencies. Most importantly, except for the period 2000–2015 and quarterly returns, the 3PRFm forecast produces negative R^2_{OS}. Therefore, Table OA.3 supports the conclusion that p-hacking could potentially be responsible for positive R^2_{OS} produced by other forecast combination techniques.

References

Clark TE, West KD (2007) Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138(1):291–311.

Harvey CR, Liu Y, Zhu H (2016) ... and the cross-section of expected returns. *Review of Financial Studies* 29(1):5–68.

Paye BS, Timmermann A (2006) Instability of return prediction models. *Journal of Empirical Finance* 13(3):274–315.

Rapach DE, Strauss JK, Zhou G (2010) Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies* 23(2):821–862.

Rapach DE, Wohar ME (2006) Structural breaks and predictive regression models of aggregate U.S. stock returns. *Journal of Financial Econometrics* 4(2):238–274.

Table OA.1
This table reports the quasi-out-of-sample performance measures of 3PRF and 3PRFm combinations of 10, 20, and 30 spurious predictors. There are $T = 276$ observations in the whole sample and $T_{OS} = 140$ observations in the quasi-out-of-sample period. "Mean(R^2_{OS})" and "$q_{0.05}(R^2_{OS})$" are the mean and 5% upper quantile of the out-of-sample R^2; Prob($R^2_{OS} < 0$) is the empirical probability of negative R^2_{OS}. "Mean(t)" and "$q_{0.05}(t)$" are the mean and 5% upper quantile of the Clark and West (2007) test statistic; "Prob($t > 1.645$)" is the empirical probability that the Clark and West (2007) test statistic exceeds 1.645.

	$A = 10$		$A = 20$		$A = 30$	
	3PRF	3PRFm	3PRF	3PRFm	3PRF	3PRFm
Mean(R^2_{OS}),%	11.24	-6.99	21.34	-14.44	29.29	-21.89
$q_{0.05}(R^2_{OS})$,%	18.70	-4.59	29.34	-10.67	37.32	-16.98
Prob($R^2_{OS} < 0$),%	1.10	100.00	0.00	100.00	0.00	100.00
Mean(t)	3.92	-2.43	5.19	-3.38	5.95	-4.07
Prob($t > 1.645$),%	99.30	0.00	100.00	0.00	100.00	0.00
$q_{0.05}(t)$	5.11	-1.23	6.30	-2.21	6.92	-2.88

Table OA.2
This table reports the quasi-out-of-sample performance measures of 3PRF and 3PRFm combinations of spurious predictors in alternative settings. In the first two columns spurious predictors are obtained using the critical value of 3 instead of 2. In the second two columns spurious predictors are correlated as described above. In the last two columns simulated returns are replaced with actual excess returns on the S&P 500 index, including dividends, in the sample from January 1947 to December 2015. There are $T = 276$ observations in each time series, and the quasi-out-of-sample period contains $T_{OS} = 140$ observations. "Mean(R^2_{OS})" and "$q_{0.05}(R^2_{OS})$" are the mean and 5% upper quantile of the out-of-sample R^2; Prob($R^2_{OS} < 0$) is the empirical probability of negative R^2_{OS}. "Mean(t)" and "$q_{0.05}(t)$" are the mean and 5% upper quantile of the Clark and West (2007) test statistic; "Prob($t > 1.645$)" is the empirical probability that the Clark and West (2007) test statistic exceeds 1.645.

	Critical value "3"		Correlated signals		Real returns	
	3PRF	3PRFm	3PRF	3PRFm	3PRF	3PRFm
Mean(R^2_{OS}),%	30.79	-12.07	3.16	-9.94	18.10	-10.46
$q_{0.05}(R^2_{OS})$,%	39.09	-8.94	7.24	-6.28	25.95	-7.68
Prob($R^2_{OS} < 0$),%	0.10	100.00	4.90	100.00	0.10	100.00
Mean(t)	5.91	-3.56	2.68	-2.97	4.31	-2.98
Prob($t > 1.645$),%	99.90	0.00	98.20	0.00	100.00	0.00
$q_{0.05}(t)$	6.79	-2.64	3.63	-1.79	5.25	-1.88

Table OA.3
This table reports R^2_{OS} and Clark and West (2007) p-values for forecasts of excess market returns constructed from 13 predictors (in Panel A for quarterly returns) or 12 predictors (in Panel B for monthly returns) using seven forecast combination techniques. The sample starts in January 1947. The left column shows the out-of-sample periods for which the statistics are computed.

		EW	MSFE	DMSFE	OLS	PCR	3PRF	3PRFm
Panel A: Quarterly returns								
1965–2005	R^2_{OS},%	3.35	3.33	3.40	-18.05	0.48	2.53	-11.23
	p-value	0.00	0.00	0.00	0.00	0.08	0.00	0.89
1965–2015	R^2_{OS},%	2.76	2.76	2.84	-24.66	0.08	2.27	-6.14
	p-value	0.00	0.00	0.00	0.04	0.09	0.00	0.48
1976–2005	R^2_{OS},%	1.12	1.06	1.01	-32.42	-3.70	-4.04	-15.50
	p-value	0.09	0.11	0.10	0.04	0.38	0.11	0.97
1976–2015	R^2_{OS},%	0.91	0.89	0.87	-37.76	-3.11	-2.59	-7.16
	p-value	0.09	0.09	0.09	0.27	0.38	0.08	0.54
2000–2005	R^2_{OS},%	3.55	3.02	3.17	-1.10	9.29	16.66	-4.03
	p-value	0.03	0.03	0.05	0.36	0.08	0.01	0.70
2000–2015	R^2_{OS},%	1.69	1.42	1.62	-30.70	2.97	7.66	6.94
	p-value	0.05	0.07	0.07	0.73	0.12	0.01	0.05
Panel B: Monthly returns								
		EW	MSFE	DMSFE	OLS	PCR	3PRF	3PRFm
1965–2005	R^2_{OS},%	1.17	1.18	1.18	-0.58	0.06	0.44	-1.44
	p-value	0.00	0.00	0.00	0.00	0.11	0.00	0.27
1965–2015	R^2_{OS},%	1.04	1.04	1.05	-1.39	-0.05	0.65	-0.99
	p-value	0.00	0.00	0.00	0.00	0.12	0.00	0.20
1976–2005	R^2_{OS},%	0.30	0.30	0.22	-2.15	-0.95	-2.34	-0.81
	p-value	0.15	0.15	0.20	0.02	0.37	0.11	0.24
1976–2015	R^2_{OS},%	0.35	0.34	0.30	-2.80	-0.84	-1.36	-0.39
	p-value	0.10	0.11	0.13	0.03	0.37	0.07	0.16
2000–2005	R^2_{OS},%	1.07	1.02	1.11	-8.77	3.78	5.78	-2.28
	p-value	0.06	0.07	0.07	0.78	0.05	0.03	0.87
2000–2015	R^2_{OS},%	0.71	0.69	0.75	-6.31	1.17	3.19	-0.36
	p-value	0.08	0.09	0.08	0.34	0.08	0.04	0.34