Comment on "Evaluating Predictive Densities of U.S. Output Growth and Inflation in a Large Macroeconomic Data Set" by Barbara Rossi and Tatevik Sekhposyan.*

Guillaume Chevillon
ESSEC Business School, CREAR

September 20, 2013

1 Introduction

The paper by Rossi and Sekhposyan (2013, RS henceforth) conduct an extensive empirical evaluation of the quality of the macroeconomic forecasting models that have most commonly been proposed. As opposed to the standard metrics which mostly assess the accuracy of point forecasts, RS is part of a recent trend of papers which consider the quality of the predictive density, i.e. the probability density function of the forecast. At a time when economic models used by Central Banks face challenges related to nonlinearities and asymmetry (witness e.g. the debate surrounding the zero lower bound on interest rates), predictive densities constitute a natural alterative to first- and second-moment based criteria for the measurement of forecast uncertainty.

Density evaluation may seem a daunting task since what is at stake is the evaluation of a complete functional form. For this reason, the literature has often focused on the evaluation of a few moments of the distribution. Yet, a convenient technique consists in using the Probability Integral Transform (PIT) of Rosenblatt (1952) applied to predictive densities by Diebold, Gunther and Tay (1998, DGT), see also Corradi and Swanson (2006). The PIT idea derives from the property that if a random variable X admits the cumulative distribution function (cdf) F_X, then $F_X(X)$ is uniformly distributed on $(0,1)$. To extend this property to conditional predictive densities, let y_{t+1} be the realization of the random variable Y_{t+1} which admits the cdf Φ_{t+1} conditional on the information set $\mathcal{I}_t$, with $t = 1, ..., T$. DGT then show that, under some regularity conditions, Φ_{t+1} evaluated at Y_{t+1} constitutes a sequence of *i.i.d.* $\mathsf{U}(0,1)$ random variables:

$$Z_{t+1} \equiv \Phi_{t+1}(Y_{t+1}|\mathcal{I}_t) = \int_{-\infty}^{Y_{t+1}} \phi_{t+1}(u|\mathcal{I}_t)\, du \overset{i.i.d}{\sim} \mathsf{U}(0,1) \tag{1}$$

*I am grateful to Laurent Ferrara and Dick van Dijk for organizing the IIF workshop on *Forecasting the Business Cycle* at the Banque de France and for editing the special issue of the International Journal of Forecasting; I also thank Barbara Rossi and Tatevik Sekhposyan for sharing their results with me.

where ϕ_{t+1} is the predictive density. Replacing densities in (1) with their model-based estimations asymptotically yields the same result under correct specification. RS assess the empirical properties of the sequence of realizations $\{z_{t+1}\}$ thus generated for various standard models (at several forecast horizons) using a battery of tests: these are designed to assess jointly or separately each aspect of the $i.i.d$ $\mathsf{U}(0,1)$ assumption for Z_{t+1}.

The analysis of Rossi and Sekhposyan shows that most models tend to reject one of the distributional assumptions of the PIT, to the exception of "meta-models" which consist in pooling others. This is in particular the case of the equally-weighted model average at a four-quarter ahead horizon and the one-quarter ahead Bayesian Model Average. This result corroborates the literature that considers forecast model combination and shows that they present better robustness properties. In particular, forecast averaging is known to exhibit robustness to misspecification and instabilities in the conditional distribution (see e.g. Clements and Hendry, 2004). This is a property shared also by direct multi-step forecasting models such as those used by RS (see Peña, 1994, and Chevillon, 2009).

This comment explores the finite sample properties of the PIT for multi-step forecast horizons and in the presence of breaks. A Monte Carlo analysis assesses distributional aspects of the PIT where a correction is proposed for the serial correlation induced by multi-step projections. A short empirical evaluation follows. The data and programs are available from the author's website.

2 Probability Integral Transform for multi-step forecasts

Whereas the $i.i.d.$ properties of the PIT for one-step ahead forecasts are well established, those for multi-step forecasts depend on the model. Denote by $q_h(\cdot)$ the density of the PIT $Z_{h,t}$ based on the model used for forecasting Y_{t+h} conditional on $\mathcal{I}_t$. For the $Z_{h,t}$ to be $i.i.d.$, the joint density $q(Z_{h,h+1}, Z_{h,h+2}, ..., Z_{h,t})$ must be factorizable into $q(Z_{h,h+1})\, q(Z_{h,h+2}) ... q(Z_{h,t})$. When $h > 1$, this is not necessarily possible. Hence DGT and Clements and Smith (2000) suggest partitioning the data into h non-overlapping subsamples (indexed by $j = 0, ..., h-1$) consisting of observations y_{j+ht}, for t such that $ht \leq T - j$. This is the approach followed by RS. For this suggestion to be feasible, the forecasting model must be conditional on the information set restricted to one subsample only, i.e. be non-overlapping. Indeed, considering e.g. the subsample defined by $j = 0$, the joint density of the PIT is given by the change of variable formula as

$$q(z_h, z_{2h}, ..., z_{kh}) = \left| \begin{array}{ccc} \frac{\partial y_h}{\partial z_h} & \cdots & \frac{\partial y_h}{\partial z_{kh}} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_{kh}}{\partial z_h} & \cdots & \frac{\partial y_{kh}}{\partial z_{kh}} \end{array} \right| \phi(y_h, ..., y_{kh}) = \frac{\partial y_h}{\partial z_h} ... \frac{\partial y_{kh}}{\partial z_{kh}} \phi(y_h, ..., y_{kh})$$

where ϕ denotes the joint density associated with the subsample. For the Z_{kh} to be $i.i.d.$, it must hold that $\phi(y_h, ..., y_{kh})$ can be factorized as

$$\phi(y_h, ..., y_{kh}) = \phi\left(y_{kh} | y_h, ..., y_{(k-1)h}\right) \phi\left(y_{(k-1)h} | y_h, ..., y_{(k-2)h}\right) ... \phi(y_h).$$

For instance, the model $y_{t+h} = \rho_h y_t + \lambda_h y_{t-h} + w_{h,t+h}$ is admissible but $y_{t+h} = \rho_h^* y_t + \lambda_h^* y_{t-1} + \omega_{h,t+h}$ is not. More generally, the Autoregressive Distributed Lag models (ADL) considered by RS should only be of the form

$$y_{t+h} = \sum_{j=0}^{J} (\alpha_j y_{t-jh} + \beta_j x_{t-jh}) + w_{h,t+h}, \quad J \geq 0,$$

for the $i.i.d$ property of the PIT to hold. These models seem restrictive and of limited empirical interest.

Under correct specification of the forecasting model, the forecast errors $w_{h,t+h} = y_{t+h} - \mathsf{E}[y_{t+h}|\mathcal{I}_t]$ is known to follow a moving average process $\mathrm{MA}(h-1)$. Hence in this comment, I consider fitting the $\mathrm{MA}(h-1)$:

$$\widehat{w}_{h,t+h} = \varepsilon_{t+h} + \theta_1 \varepsilon_{t+h-1} + ... + \theta_{h-1} \varepsilon_{t+1} \tag{2}$$

where $\widehat{w}_{h,t+h}$ denotes the observed multi-step forecast errors.[1] I compare the properties of the PIT of Y_{t+h} (equivalently $\widehat{w}_{h,t+h}$) to those based on the residual $\widehat{\varepsilon}_{t+h}$ from the estimation of (2).[2]

3 Monte Carlo

3.1 Set up

I follow the set-up of RS and consider the data generating process (DGP):

$$y_{t+1} = a(L) y_t + x_t + \varepsilon_{t+1}, \quad \varepsilon_t \overset{i.i.d}{\sim} \mathsf{N}(0,1)$$

for $t = 1, ...T = 300$ and where $a(L)$ is a lag polynomial of order $p \leq 1$ (namely model 1: $a(L) = .7$ or model 2: $a(L) = .2 + .5L$) and x_t is generated as

$$x_t = x_{t-1} + \sigma_t \nu_t$$
$$\sigma_t \overset{i.i.d}{\sim} \mathsf{N}(0,2), \quad \nu_t \overset{i.i.d}{\sim} \mathsf{Bernoulli}(.05)$$

The choice of forcing variable x_t is made to allow for possible shifts in its conditional distribution that may lead to misspecification in the estimated model as in RS. Figure 1 presents one realization of (y_t, x_t) where $a(L) = 0.2 + 0.5L$. The parameter of the Bernoulli distribution is set sufficiently high so some shifts are likely to occur towards the end of the sample. Yet the variance of σ_t is not so large as to induce significant nonstationarity (when tested as the presence of a unit root). The type of break is made to reflect those that have been shown to favor the pooling of forecasting models.

The forecasting model is the ADL, for $h \in \{1, 4\}$:

$$y_{t+h} = \mu + \alpha(L) y_t + \beta(L) \widetilde{x}_t + w_{h,t+h} \tag{3}$$

[1] For simplicity and comparability, the estimated variances of the residuals do not correct for the estimation of parameters.

[2] Throughout, I estimate Moving Averages by nonlinear least-squares.

where the orders $p \leq p_{\max}$, $q \leq q_{\max}$ of the lag polynomials $\alpha(L), \beta(L)$ are chosen by the BIC as in RS. The model is estimated over $P = 100$ rolling windows of $R = 200$ observations. The regressor $\widetilde{x}_t$ is a step dummy that takes value 1 over the latest $q_{\max}$ observations of the rolling sample and zero beforehand. Hence the estimated model aims to capture the most recent shift (if any occurs at the very end of the sample). I set $p_{\max} = 4$ and $q_{\max} = 5$ and use 1,000 Monte Carlo replications.[3]

3.2 Evaluation

I assess some properties of the PIT derived from model (3) under the assumption of normality of the multi-step forecast errors $\widehat{w}_{h,t+h}$. I compare them to those based on the residuals $\widehat{\varepsilon}_t$ obtained from fitting an MA$(h-1)$ to $\widehat{w}_{h,t+h}$. I assess the PIT via three of the tests proposed by RS: (a) a Kolmogorov-Smirnov test KS for uniform distribution of the PIT; (b) Ljung-Box tests Q and Q2 for the absence of autocorrelation in the levels and squares; and (c) the Doornik-Hansen (DH) test of normality for the PIT that have been transformed via the Gaussian inverse cdf. For each test, I report the Monte Carlo average test statistic, with its associated asymptotic p-value (conclusions are similar when considering the median p-value over the Monte Carlo replications).

Table 1 presents results from the simulation based on the two models where one or two lags of y_t enter on the right-hand side in the DGP. Both models yields similar results here (although p-values are larger for Model 2). The specification test based on the KS statistic does not reject for $h = 1$ but does so when considering the fourth-horizon multi-step residuals. The MA$(h-1)$ correction performs well here as the null of correct specification does not reject at conventional sizes. The Ljung-Box statistics show that latter correction appears to work by reducing the serial correlation of the PIT. Whereas the Ljung-Box Q and Q2 tests do not reject at horizon $h = 1$, they do when considering the multi-step residuals. The corrected multi-step residuals present less autocorrelation although Q still rejects despite an improvement in the associated p-value. Finally, the normality test does not reject in any case.

4 Empirical Analysis

I now estimate model (3) and perform the above tests using quarterly data ranging from 1947(1) to 2013(1) obtained from the FRED database maintained by the Federal Reserve of Saint Louis.[4] We consider forecasting the quarterly changes in log real GDP and log real per capita GDP, as well as the quarterly changes in inflation measured as the quarterly change in log CPI, log core CPI or log GDP deflator. The PIT evaluation output is collected in table 2. In addition to RS, I also forecast the actual inflation level (in table 3). I set $P = 80$ so $R = 184$ or 183 depending on whether the variable of interest is the first or second difference in the original data.

[3] All simulations and empirical results are obtained using Ox 6.30, see Doornik (2009). The code is available from my website.

[4] The data on core CPI starts in 1957(1). The data is available from my website.

The Kolmogorov-Smirnov test of correct specification for the predictive density shows least rejection when performed for the corrected $h = 4$ multi-step residuals. Yet, it rejects at the 5% size for real GDP growth and core inflation (level and change). The non corrected residuals lead to rejection for all variables but the change in CPI inflation. At horizon $h = 1$, the KS test does not reject for changes in CPI and GDP deflator inflations.

The Ljung-Box tests for autocorrelation (up to the fourth lag) show that multi-step correction can improve the inference, especially in the squares of the PIT, yet the evidence is not strong enough to show it is enough. Interestingly RS find that the tests reject more often for the squares of the PIT, hence the multi-step correction might be helpful in correcting for autoregressive conditional heteroscedasticity.

Finally, the Doornik-Hansen specification test shows an improvement in the specification when correcting the multi-step residuals (except when considering the GDP deflator and the change in CPI inflation). It rejects less often than using a one-step ahead forecasting model.

5 Conclusions

This comment has considered one method for correcting the tests for the correct specification of the multi-step predictive density. Indeed, when testing forecasts at horizon $h > 1$ using Probability Integral Transforms, the resulting PIT are generally not independent. Recommendations found in the literature rely either (i) on the joint modelling of the forecasts at several horizons (Clements and Smith, 2000) or (ii) the joint use of non-overlapping data using Bonferroni corrections (Rossi and Sekhposyan, 2013). The latter technique is appealing in its simplicity, yet it requires not only that the tests be performed over non overlapping subsamples but that the conditioning information sets entering the forecasting models themselves be non-overlapping. The latter condition may be too strict in empirical work. Hence I assessed here a simple alternative solution which consists in fitting an MA$(h-1)$ model to the forecast errors. The PIT are then based on the residuals from this auxiliary model. This technique is intuitively appealing since it avoids dealing with non-overlapping subsamples. I show in simple simulation and empirical exercises that the proposed technique seem to improve the fit of the predictive densities.

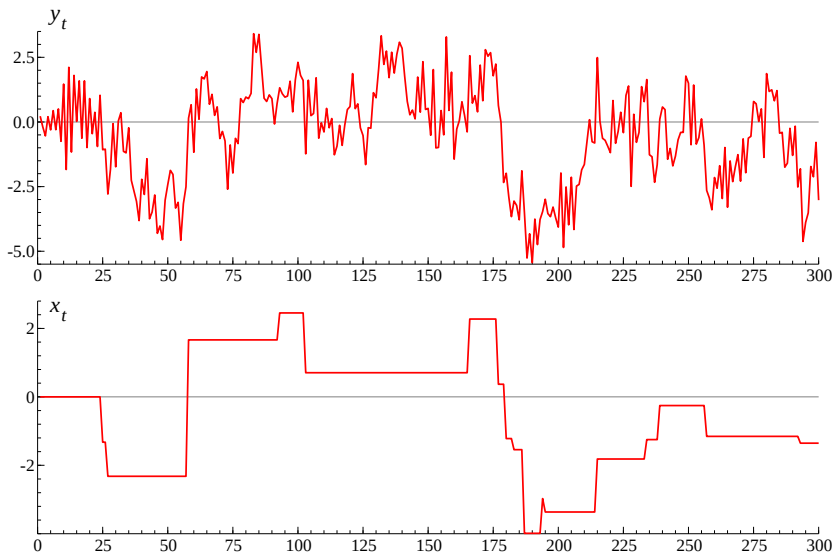

Figure 1: One realization of the processes $y_t = .2y_{t-1} + .5y_{t-2} + x_t + \varepsilon_t$ where ε_t is standard Gaussian white noise and x_t is a level shifting random variable.

	Model 1: $a(L) = .7$			Model 2: $a(L) = .2 + .5L$		
	$h = 1$	$h = 4$		$h = 1$	$h = 4$	
	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$
Kolmogorov-Smirnov						
KS statistic	1.06	1.52	0.88	0.97	1.35	0.73
p-value	0.208	0.020	0.425	0.308	0.053	0.666
Ljung-Box						
Q statistic	4.48	55.5	15.7	4.43	31.7	13.1
p-value	0.350	0.000	0.003	0.351	0.000	0.011
Q2 statistic	4.08	17.2	5.15	4.14	7.69	4.89
p-value	0.400	0.002	0.272	0.389	0.104	0.299
Doornik-Hansen						
DH statistic	2.0	2.3	1.8	2.06	1.78	1.75
p-value	0.367	0.316	0.408	0.356	0.410	0.417

Table 1: Statistics for the Probability Integral Transforms (PIT) of simulated data. Model 1 is $y_{t+1} = .7y_t + x_t + \varepsilon_{t+1}$ and Model 2 is $y_{t+1} = .2y_t + .5y_{t-1} + x_t + \varepsilon_{t+1}$. x_t is a variable that undergoes occasional level shifts. The estimated model is an ADL $y_{t+h} = \mu + \alpha(L)\, y_t + \beta(L)\, \widetilde{x}_t + w_{h,t+h}$ whose order is chosen by the BIC. $\widetilde{x}_t$ denotes a step dummy. The PIT is based on the residuals $\widehat{\varepsilon}_t$ when $h = 1$. When $h = 4$, it is either based on the multi-step residuals $\widehat{w}_{h,t+h}$ or from the residuals $(\widehat{\varepsilon}_t)$ obtained from fitting an MA$(h-1)$ to $\widehat{w}_{h,t+h}$. The number of Monte Carlo replications is 1,000. Estimation is carried over 100 rolling samples of 200 observations each.

	GDP growth			Per Capita GDP growth			$\Delta\pi^{CPI}$			$\Delta\pi^{core}$			$\Delta\pi^{GDP}$		
	$h=1$	$h=4$		$h=1$	$h=4$		$h=1$	$h=4$		$h=1$	$h=4$		$h=1$	$h=4$	
	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$
Kolmogorov-Smirnov															
KS statistic	1.68	1.82	1.66	1.51	1.56	1.25	1.06	1.06	1.07	2.55	2.57	2.33	1.33	1.36	1.33
p-value	0.007	0.003	0.008	0.022	0.015	0.088	0.212	0.209	0.206	0.000	0.000	0.000	0.059	0.049	0.057
Ljung-Box															
Q statistic	10.7	19.3	11.6	6.6	32.2	12.3	4.5	14.6	6.4	13.1	24.2	9.7	5.3	13.0	3.1
p-value	0.005	0.000	0.003	0.038	0.000	0.002	0.105	0.001	0.041	0.001	0.000	0.008	0.069	0.001	0.207
Q2 statistic	5.7	4.8	5.2	7.6	10.2	9.7	6.9	14.9	7.1	8.8	7.0	3.2	6.2	2.8	3.6
p-value	0.057	0.090	0.075	0.022	0.006	0.008	0.032	0.001	0.028	0.012	0.031	0.204	0.045	0.244	0.163
Doornik-Hansen															
DH statistic	2.6	11.2	2.4	19.2	11.0	2.8	18.9	14.3	14.7	8.3	4.1	6.6	8.0	2.2	2.6
p-value	0.270	0.004	0.308	0.000	0.004	0.244	0.000	0.001	0.001	0.016	0.127	0.038	0.018	0.332	0.266

Table 2: Statistics for the Probability Integral Transforms (PIT) of empirical data. The estimated model is an ADL $y_{t+h} = \mu + \alpha(L)\, y_t + \beta(L)\, \widetilde{x}_t + w_{h,t+h}$ whose order is chosen by the BIC. $\widetilde{x}_t$ denotes a step dummy. The PIT is based on the residuals $\widehat{\varepsilon}_t$ when $h=1$. When $h=4$, it is either based on the multi-step residuals $\widehat{w}_{h,t+h}$ or from the residuals $(\widehat{\varepsilon}_t)$ from fitting an MA$(h-1)$ to $\widehat{w}_{h,t+h}$. The data were obtained from the FRED database maintained by the Saint-Louis Fed. The quarterly variables to be forecast are real GDP growth and real per capita GDP growth, as well as change in inflation based on the consumer price index ($\Delta\pi^{CPI}$), the core CPI ($\Delta\pi^{core}$) and the GDP deflator ($\Delta\pi^{GDP}$). The tests that are reported are the Kolmogorov Smirnov test for the null that the PIT are uniformly distributed; the Q and Q2 Ljung-Box test for autocorrelation up to the fourth order lag for the level and squares of the PIT respectively; and the Doornik-Hansen test for normality of the PIT transformed via the Gaussian inverse cdf.

	π^{CPI}			π^{core}			π^{GDP}		
	$h=1$	$h=4$		$h=1$	$h=4$		$h=1$	$h=4$	
	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$	$\widehat{\varepsilon}_t$	$\widehat{w}_{h,t+h}$	$\widehat{\varepsilon}_t$
Kolmogorov-Smirnov									
KS statistic	1.42	2.16	1.17	3.13	4.06	2.60	1.72	2.04	0.89
p-value	0.036	0.000	0.131	0.000	0.000	0.000	0.005	0.000	0.404
Ljung-Box									
Q statistic	2.7	8.7	11.9	14.8	37.0	21.1	6.0	11.7	22.7
p-value	0.253	0.013	0.003	0.001	0.000	0.000	0.050	0.003	0.000
Q2 statistic	6.3	1.2	7.2	8.8	43.5	7.8	4.1	3.6	19.5
p-value	0.043	0.551	0.027	0.012	0.000	0.020	0.131	0.167	0.000
Doornik-Hansen									
DH statistic	16.0	10.8	9.5	3.1	4.2	0.5	11.3	1.0	5.4
p-value	0.000	0.005	0.009	0.216	0.125	0.761	0.004	0.603	0.068

Table 3: Statistics for the Probability Integral Transforms (PIT) of empirical data. The estimated model is an ADL $y_{t+h} = \mu + \alpha(L)\, y_t + \beta(L)\, \widetilde{x}_t + w_{h,t+h}$ whose order is chosen by the BIC. $\widetilde{x}_t$ denotes a step dummy. The PIT is based on the residuals $\widehat{\varepsilon}_t$ when $h=1$. When $h=4$, it is either based on the multi-step residuals $\widehat{w}_{h,t+h}$ or from the residuals $(\widehat{\varepsilon}_t)$ from fitting an MA$(h-1)$ to $\widehat{w}_{h,t+h}$. The data were obtained from the FRED database maintained by the Saint-Louis Fed. The variables to be forecast are quartely measures of inflation based on the consumer price index (π^{CPI}), the core CPI (π^{core}) and the GDP deflator (π^{GDP}). The tests that are reported are the Kolmogorov Smirnov test for the null that the PIT are uniformly distributed; the Q and Q2 Ljung-Box test for autocorrelation up to the fourth order lag for the level and squares of the PIT respectively; and the Doornik-Hansen test for normality of the PIT transformed via the Gaussian inverse cdf.

References

Chevillon, G. (2009). *Multi-step Forecasting in Emerging Economies: an Investigation of the South African GDP,* International Journal of Forecasting, 25(3), 602–28.

Clements, M.P. and J. Smith (2000), *Evaluating the Forecast Densities of Linear and Non-linear Models: Applications to Output Growth and Unemployment*, Journal of Forecasting, 19(4), 255–276.

Corradi, V. and N.R. Swanson (2006), Predictive Density Evaluation, in: G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol. 1, North Holland: Elsevier, 197–284.

Diebold, F.X., T.A. Gunther, and A.S. Tay (1998), *Evaluating Density Forecasts with Applications to Financial Risk Management*, International Economic Review, 39(4), 863–883.

Doornik, J. (2009). An Introduction to OxMetrics 6. London: Timberlake Consultants Press.

Doornik, J.A. and H. Hansen (2008). *An Omnibus Test for Univariate and Multivariate Normality*, Oxford Bulletin of Economics and Statistics, 70(S1), 927–939.

Hendry, D.F. and Michael P. Clements, (2004). *Pooling of forecasts*, Econometrics Journal, 7(1), pp. 1–31.

Peña, D. (1994). Discussion: Second-generation time-series model, A comment to "Some advances in non-linear and adaptive modelling in time-series analysis" by G.C. Tiao and R. S. Tsay. Journal of Forecasting 13, pp. 133–140.

Rosenblatt, M. (1952). *Remarks on a Multivariate Transformation*, Annals of Mathematical Statistics, 23(3), 470–2.