Communications in Applied Mathematics and Computational Science

ON A COMPUTATIONALLY SCALABLE SPARSE FORMULATION OF THE MULTIDIMENSIONAL AND NONSTATIONARY MAXIMUM ENTROPY PRINCIPLE

Illia Horenko, Ganna Marchenko and Patrick Gagliardini

vol. 15 no. 2 2020

msp

COMM. APP. MATH. AND COMP. SCI.
Vol. 15, No. 2, 2020
https://doi.org/10.2140/camcos.2020.15.129

ON A COMPUTATIONALLY SCALABLE SPARSE FORMULATION OF THE MULTIDIMENSIONAL AND NONSTATIONARY MAXIMUM ENTROPY PRINCIPLE

ILLIA HORENKO, GANNA MARCHENKO AND PATRICK GAGLIARDINI

Data-driven modeling and computational predictions based on the maximum entropy principle (MaxEnt principle) aim to find as simple as possible — but not simpler than necessary — models that allow one to avoid the data-overfitting problem. We derive a multivariate nonparametric and nonstationary formulation of the MaxEnt principle and show that its solution can be approximated through a numerical maximization of the sparse constrained optimization problem with regularization. Application of the resulting algorithm to popular financial benchmarks reveals memoryless models allowing for simple and qualitative descriptions of data of the major stock market indices. We compare the obtained MaxEnt models to the heteroscedastic models from computational econometrics (GARCH, GARCH-GJR, MS-GARCH, and GARCH-PML4) in terms of the model fit, complexity, and prediction quality. We compare the resulting model log-likelihoods, the values of the Bayesian information criterion, posterior model probabilities, the quality of the data autocorrelation function fits, as well as the value-at-risk prediction quality. We show that all of the seven considered major financial benchmark time series (DJI, SPX, FTSE, STOXX, SMI, HSI, and N225) are better described by conditionally memoryless MaxEnt models with nonstationary regime-switching than by the common econometric models with finite memory. This analysis also reveals a sparse network of statistically significant temporal relations for the positive and negative latent variance changes among different markets. The code is provided for open access.

The maximum entropy principle (MaxEnt principle) was originally introduced in physics and information theory to search for least-biased probabilistic data descriptions that match certain statistical properties of the data, e.g., the data distribution moments [18]. The MaxEnt principle implies that the most unbiased distribution for the data is the one that admits the most uncertainty, measured in terms of entropy. Depending on the constraints imposed in the entropy maximization problem,

The authors thank Michael Rockinger (HEC Lausanne) for providing the code for the PML4 parameter estimator we used for comparison. This work was funded by the SNSF (project 140829) and DFG SPP 1114 (Mercator Fellowship of Horenko).
MSC2010: 28D20, 37M10, 37M25, 91B84.
Keywords: machine learning, financial time series, maximum entropy, heteroscedasticity, sparsity.

different parametric probability distributions that can be described by this principle are Gaussian, exponential, Laplace, Cauchy, chi-squared, and gamma distributions, among others. This MaxEnt-based probabilistic modeling approach has been successfully applied to many problems ranging from biology [26; 24] and natural language processing [3; 25] to applications from economics and finance [31; 30].

In contrast to the parametric MaxEnt modeling, where the particular parametric distribution models (Gaussian, exponential, etc.) are dependent on the fixed finite set of constant parameters, computation of nonparametric MaxEnt densities even in one dimension is not a trivial task, as was previously discussed in [1; 23; 15]. In [22] a systematic mathematical derivation for a nonstationary extension of the nonparametric MaxEnt methodology for one-dimensional time series problems was introduced. This BV-entropy framework imposes a mild bounded-variation (BV) assumption on the time dependence of the nonstationary moments of the underlying nonstationary probability distribution function (p.d.f.). In [22] it was shown that the original ill posed nonstationary MaxEnt principle — formulated as an entropy maximization problem with time-dependent moment constraints — can be sharply bounded from below via a well posed and computationally scalable maximization problem. This lower-bound problem appeared to be a nonparametric regime-switching entropy maximization problem with K locally stationary regimes, subject to K l_1-constraints on the regime-specific vectors of moments and to a BV-constraint for the latent regime transitions. The linear BV-constraint happens to bound from above with C the maximum number of regime switches — and controls the persistence of the obtained MaxEnt regime transition models. It was also shown that the optimal values of C and K can be estimated by deploying common model selection criteria like the Bayesian information criterion (BIC).

However, this model formulation is confined to one-dimensional time series analysis problems only. A direct extension of this one-dimensional BV-entropy methodology to multiple dimensions is hampered by the curse of dimension and the overfitting problems: linear growth in the number of problem dimensions will result in the exponential growth in the number of underlying multivariate MaxEnt parameters that have to be determined from the data statistics. In many practical applications — for example in economics and finance — there is only one historical realization of the process that is available, with no possibility to obtain other sampling realizations from some "model". For every particular dimension at every particular time there is only one data point available. This means that a completely nonparametric approach to statistical analysis of such data leads to ill posed problems [13].

In the following, we describe a computational algorithm that achieves a sparse multidimensional extension of the one-dimensional BV-entropy model from [22], where the coupling between the n individual one-dimensional entropy maximization

problems will be achieved through a sparsifying regularization constraint that controls appropriate function space distances (l_2, l_1, or BV) between the individual latent factors.

The resulting algorithm (further referred to as TV-entropy) allows a computationally tractable search for the most unbiased multivariate time-dependent distribution of the data, minimizing the optimal number of locally stationary hidden regimes and sparsifying their time-persistent regime-switching dynamic and the regime-specific distribution parameter vectors. Then we illustrate its performance and compare it to common approaches on the test model system (a regime-switching Gaussian, Figure 1). Finally, we apply this algorithm and the common GARCH tools to a set of seven popular financial stock market indices and show that these simpler memoryless MaxEnt models of volatility outperform the popular heteroscedastic methods with finite memory for all benchmarks considered (see Figures 2–4 and Table 2).

Method

We start with n-dimensional multivariate time series data $x_{t,i}$ for all $t \in \{1, \ldots, T\}$ and all $i \in \{1, \ldots, n\}$ on a closed interval $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \subset \mathbb{R}^n$. The goal is to find the most descriptive and least complex model for this data. We will first assume that x_t is conditionally independent in time, and would like to estimate its unknown marginal time-dependent probability densities $f_{t,i}(x)$. In the context of the MaxEnt principle, the density $f_{t,i}(x)$ can be identified by solving an optimization problem which consists of finding a time-dependent density function with the maximum entropy among all distribution functions that match the data in the first $m + 1$ sample moments at all of the given time instances t in the dimension i. There is a long tradition in statistics and econometrics of adopting inference methodologies based on matching sample moments (generalized method of moments (GMM) estimation [12]). Moreover, MaxEnt methods are often invoked in multinomial choice problems [11; 21]. Maximization of the expected entropy with respect to a time-dependent multivariate probability density $f_{t,i}(x)$ (where the expectation is taken over the time t and the dimension n) can be written as

$$\max_{f_{t,i}(x) \text{ for all } t,\, i} \left\{ \mathbb{E}_{t,i}[H[f_{t,i}(x)]] = \mathbb{E}_{t,i}\left[-\int_{\mathcal{X}_i} f_{t,i}(x) \ln f_{t,i}(x)\, dx \right] \right\}, \tag{1}$$

subject to

$$\int_{\mathcal{X}_i} x^j f_{t,i}(x)\, dx = \mu_{t,i}(j) \quad \text{for all } j \in \{0, \ldots, m\},\ i \in \{1, \ldots, n\}, \text{ and } t \in \{1, \ldots, T\}, \tag{2}$$

and $\mu_{t,i} \in \mathcal{R}^{m+1}$ are time- and dimension-dependent sample moments. Then the optimal $f^*_{t,i}(x)$ can be derived by computing first-order optimality conditions of

the corresponding optimization problem (1)–(2), providing the formulation

$$f_{t,i}(x) = \exp\left[-\sum_{j=0}^{m} \Lambda_{t,i}(j)x^j\right] \quad \text{for all } t, i \tag{3}$$

such that

$$\int_{\mathscr{X}_i} x^n \exp\left[-\sum_{j=0}^{m} \Lambda_{t,i}(j)x^j\right] dx = \mu_{t,i}(n) \quad \text{for all } n \in \{0, \dots, m\}, \tag{4}$$

$$\mu_{t,i}(0) = 1, \tag{5}$$

where $\Lambda_{t,i} \in \mathcal{R}^{m+1}$ are unknown time-dependent parameters (Lagrange multipliers) of MaxEnt distributions. One way to compute $\Lambda(\cdot)$ would be to maximize the log-likelihood function based on the obtained densities $f_{t,i}(x)$, i.e.,

$$\max_{\Lambda(\cdot)} \mathscr{L}(\Lambda(\cdot)) = -\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\sum_{j=1}^{m} \Lambda_{t,i}(j)x_t^j + \ln Z_{\Lambda_{t,i}}\right), \tag{6}$$

$$\text{where} \quad Z_{\Lambda_{t,i}} = \int_{\mathscr{X}_i} \exp\left[-\sum_{j=1}^{m} \Lambda_{t,i}(j)x^j\right] dx = \exp[\Lambda_{t,i}(0)]. \tag{7}$$

The first-order optimality conditions of the problem (6) are equivalent to conditions (3)–(5); therefore, $\Lambda(\cdot)$ maximizing the criterion in (6) are the optimal values of the MaxEnt densities parameters in (3)–(5).

In realistic applications when only one historical realization sequence $\{x_1, \dots, x_T\}$ is available, there is no straightforward solution to the nonstationary problem (6), as at each time instance t we have many more parameters than we have observed data. To solve this problem, we are going to introduce the two following assumptions.

Assumption 1. Total variation (a TV-norm) of the MaxEnt parameters $\Lambda(\cdot)$ is bounded; i.e.,

$$|\Lambda|_{\mathrm{TV}} = \sum_{t_1,t_2=1}^{T}\sum_{i_1,i_2=1}^{n} |\Lambda_{t_1,i_1}(\cdot) - \Lambda_{t_2,i_2}(\cdot)|_1 = C < +\infty. \tag{8}$$

Assumption 2. There exist $K \ll nT$ distinct sets of parameters $\lambda^{(k)}$, $k = 1, \dots, K$, and $\gamma_{t,i} = [\gamma_{t,i}^{(1)}, \dots, \gamma_{t,i}^{(K)}]$ (with $\sum_{k=1}^{K} \gamma_{t,i}^{(k)} = 1$ and $\gamma_{t,i}^{(K)} \geq 0$ for all t, i, k), such that for any t, i, and k the vector $\Lambda_{t,i}$ can be expressed as a convex linear combination

$$\Lambda_{t,i} = \sum_{k=1}^{K} \gamma_{t,i}^{(k)} \lambda^{(k)}. \tag{9}$$

Assumptions 1 and 2 introduce sparsity in $\Lambda_{t,i}$ across time and space indices t and i. Very importantly, these assumptions do not rely on any ordering in the space dimension. Indeed, in many practical applications, e.g., in economics and finance,

a natural ordering across assets, institutions, markets, etc., does not exist. On a practical side, for real financial data these assumptions will be fulfilled automatically if one sets both constants K and C to be large enough (for example, setting $K = nT$ always fulfills Assumption 2). In the practical applications the aim will be in finding the computationally scalable lower-bound estimates of these constants. Substituting condition (9) in (8), using Jensen's inequality, and inserting into (6) the obtained inequality constraint as the penalty term of the Karush–Kuhn–Tucker conditions leads to the following lower-bound approximation of the MaxEnt problem (6):

$$\max_{\lambda,\gamma} L(\lambda,\gamma) = -\sum_{k=1}^{\boldsymbol{K}}\Bigg[\sum_{i=1,\,t=1}^{n,T}\gamma_{t,i}^{(k)}\Bigg(\sum_{j=1}^{m}\lambda_j^{(k)}x_{t,i}^{j}+Z_i^{(k)}\Bigg) + \sigma_C|\lambda^{(k)}|_1\sum_{t_1,t_2=1}^{T}\sum_{i_1,i_2=1}^{n}|\gamma_{t_1,i_1}^{(k)}-\gamma_{t_2,i_2}^{(k)}|\Bigg] \quad (10)$$

such that

$$Z_i^{(k)} = \ln\int_{\mathscr{X}_i}\exp\Bigg[-\sum_{j=1}^{m}\lambda_{(j)}^{(k)}x^j\Bigg]dx, \quad \gamma_{t,i}^{(k)} \geq 0, \quad \sum_{k=1}^{\boldsymbol{K}}\gamma_{t,i}^{(k)} = 1, \quad \sigma_C \geq 0. \quad (11)$$

A solution of the obtained lower-bound problem is an approximation to the solution of the original nonstationary ill posed MaxEnt problem (1). The optimization criterion in (10) is nonlinear and nonconvex — implying that the problem can have more than one locally optimal solution. However, there are three properties of this optimization problem formulation that can be exploited numerically: (i) when λ is kept fixed, (10)–(11) becomes a uniquely solvable linear programming (LP) problem with respect to γ — and can be solved very efficiently by means of common LP-tools (e.g., with the simplex method); (ii) if $\sigma_C = 0$, solving (10)–(11) becomes equivalent to solving n independent one-dimensional nonstationary entropy maximization problems from the BV-entropy method in [22]; (iii) the algebraic structure of the term containing σ_C is similar to the LASSO-regularization formulation very popular in machine learning [29], and increasing σ_C will result in increasing the sparsity of λ and in penalizing the temporal and cross-sectional variations in γ — making the obtained MaxEnt model approximations more simple, sparse, and persistent across dimensions and in time. For a numerical solution of the problem (10)–(11) with a fixed set of parameters K and σ_C we will adapt the subspace algorithms introduced in [16; 17] (that were further developed in [9; 27]): to get the best possible use of the algebraic structure in (10)–(11), resulting algorithms should iterate between two distinct optimization problems, where the problem (10)–(11) is solved with respect to Λ (for current fixed values of Γ) and in the following step with respect to Γ (for current fixed values of Λ). The Λ-optimization step will involve the independent solution of $\boldsymbol{K}$ regularized stationary MaxEnt problems. As a result, an iterative

procedure would converge monotonically to a local maximum solution of the problem (10)–(11). This procedure — further referred to as TV-entropy — should be repeated for different combinations of the input parameters $\boldsymbol{K}, \sigma_C$ (from some predefined discrete sets), and common model discrimination tools like the Akaike information criterion [6], cross-validation [19], or bootstrap [6; 19] will be deployed to identify the most optimal combination of — hopefully small — parameters $\boldsymbol{K}, \sigma_C$.

It is straightforward to verify that the second term in (10) can alternatively be formulated as a set of two linear inequality constraints

$$|\lambda^{(k)}|_1 \leq C_{\lambda^{(k)}} \quad \text{for all } k = 1, \ldots, \boldsymbol{K}, \tag{12}$$

$$\sum_{t_1,t_2=1}^{T} \sum_{i_1,i_2=1}^{n} |\gamma_{t_1,i_1}^{(k)} - \gamma_{t_2,i_2}^{(k)}| \leq C_\gamma \quad \text{for all } k = 1, \ldots, \boldsymbol{K}, \tag{13}$$

where $C_\gamma \sum_{k=1}^{\boldsymbol{K}} C_{\lambda^{(k)}} \leq C$, with C being the global TV-constant defined in (8). This formulation allows a separate handling of constraints for the regime-switching and for the MaxEnt parameters with two separate sets of constraining variables $C_{\lambda^{(k)}}$ and C_γ. In contrast, the Tykhonov-like formulation (10)–(11) allows joint simultaneous control for all of the variables by means of a single constraining variable σ_C.

Results

First, we illustrate an application of the TV-entropy methodology introduced above and its comparison to the common regime identification methods like HMM and the adaptive Gaussian moving window (GMW) in a study with artificial simulated data. For this simulated data study we use the model system proposed in [22]: we generate multiple samples from the regime-switching model with two Gaussian regimes, i.e., $\tilde{x}_t = \sum_{i=1}^{2} \gamma_t^{(i)} \tilde{x}_t^{(i)}$, where $\tilde{x}_t^{(i)} \sim \mathcal{N}(0, v^{(i)})$ with $v^{(1)} = 1$ and $v^{(2)} = \tilde{v}$. For each value of $v^{(2)}$ we generate 100 samples with 1000 points. The regime-switching weights are discrete and satisfy the convexity constraints in (11), imposing only one active regime at time t. There are three regime transitions in the data-generating process that occur every 250 points. The time-dependent variance signal can then be computed as $v_t = \sum_{i=1}^{2} \gamma_t^{(i)} v^{(i)}$. This test system is designed to be favorable for HMM and GMW — since their common variants rely on the Gaussianity assumption for the realizations. The TV-entropy models were estimated with two regimes, six density regime parameters, and between one and ten regime switches. We used ten annealing steps for estimating both the TV-entropy and HMM models. The corresponding optimal parameters were chosen using BIC. For the GMW model we used bandwidth value $b = \{10, 30, 50\}$ to obtain various levels of persistence. As demonstrated in Figure 1, TV-entropy outperforms these common models in data classification and reconstruction of the variance signal used in the data-generating process.

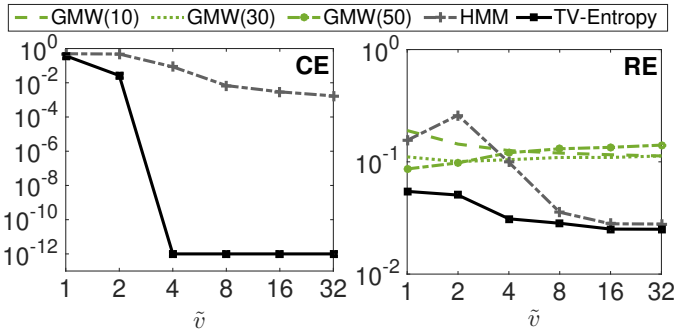

Figure 1. Average classification error (CE) between the true and reconstructed switching regime paths (left) and average relative error (RE) between the true and the reconstructed variance signals (right) in log scale obtained for 100 two-regime Gaussian samples with $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \tilde{v})$ regimes.

	stationary, without regime transitions	nonstationary, with regime transitions
parametric	GARCH/GARCH-GJR	MS-GARCH
semi- or nonparametric	GARCH-PML4	TV-entropy

Table 1. Classification of the volatility models used in the comparison.

In the following empirical study we compare the performance of the TV-entropy approach (10)–(11) to different popular variants of GARCH models (Table 1). For the sake of simplicity, we will start with $\sigma_C = 0$. For a comparison with common econometric models we use a classic GARCH[1] model with normally distributed innovations [4] and a GARCH-PML4 model that uses the MaxEnt principle to achieve the less restrictive description of the density [15]. Additionally, we employ the GARCH-GJR model that incorporates the asymmetric influence of positive and negative news on volatility [10] and the MS-GARCH model that assumes the presence of several hidden GARCH regimes in data with the regime transitions governed by a Markov chain [2].

As benchmark problems we considered the daily percentage log-returns time series of the seven major world market indices: DJI, SPX, FTSE, STOXX, SMI, HSI, and N225.[2] The data is available in the Oxford-Man Institute's "Realized

[1] With this abbreviation we refer to the GARCH(1, 1) model.

[2] DJI (Dow Jones Industrial Average Index, United States), SPX (Standard & Poor's 500 Index, United States), FTSE (Financial Times Stock Exchange 100 Index, United Kingdom), STOXX (EURO STOXX 50 Index, European Union), SMI (Swiss Market Index, Switzerland), HSI (Hang Seng Index, Honk Kong and China), N225 (Nikkei Stock Average Index, Japan).

index	T	skewness	kurtosis
DJI	2864	0.08	10.96
SPX	2862	−0.08	10.11
FTSE	2878	−0.09	6.81
STOXX	2896	−0.12	7.61
SMI	2875	0.03	9.26
HSI	2603	0.13	16.38
N225	2773	−0.41	13.30

Table 2. Description of the data samples, where T is a sample size.

K	C_γ	N	k_i	$C_{\lambda^{(i)}}$
$\{1, 2, 3, 4\}$	$\{1, 2, 3, \ldots, 50\}$	10	6	$+\infty$

Table 3. Input parameters, where K is the number of hidden regimes, C_γ the maximum number of transitions per regime, N the number of annealing steps used during estimation, k_i a number of moment constraints in the MaxEnt regime i, and $C_{\lambda^{(i)}}$ the l_1 bound.

Library" [14].[3] The sample size and some related statistics of all considered samples are gathered in Table 2. All considered samples are not normally distributed.[4] High kurtosis points to the presence of the fat tails in data, and nonzero skewness suggests asymmetry in the distribution of the returns. These properties are generally consistent with empirical data.

To reduce the number of estimation routines needed to obtain the optimal parameters of the TV-entropy, we split the estimation procedure into two stages. First, models are estimated for all combinations of the input parameters outlined in Table 3 without l_1-regularization of the regime parameters. The optimal number of regimes (K^*) and transition upper bounds (C_γ^*) for each data set are chosen based on minimal BIC value.

Next, in order to identify the optimal number of local parameters needed to describe data in each of the regimes, we estimate TV-entropy models with various values of l_1-regularization bounds, while keeping the number of hidden regimes and transitions fixed according to their optimal values (K^*, C_γ^*) obtained at the previous step. The range of l_1 bounds $C_{\lambda^{(i)}}$ is data-dependent and varies across samples. To choose the appropriate range for each sample we analyze the corresponding unregularized solution.

[3] See https://github.com/Ganna85/TV-Entropy for the data set and source code used in numerical experiments.

[4] For the normal distribution the skewness and kurtosis should be equal to zero and four, respectively.

index	K^*	C_γ^*	k^*
DJI	3	21	[6, 6, 4]
SPX	3	18	[6, 6, 4]
FTSE	4	12	[6, 6, 6, 6]
STOXX	3	16	[6, 6, 4]
SMI	3	19	[6, 6, 4]
HSI	3	8	[6, 4, 4]
N225	3	11	[6, 6, 4]

Table 4. Best regularized TV-entropy models, where K^*, C_γ^* are the optimal values of parameters with respect to the BIC and k^* is the vectors with the optimal number of regime parameters.

As shown in Table 4, the application of the TV-entropy model reveals three hidden regimes in all of the considered benchmarks, with an exception of the FTSE index data, where the four-regime model is identified as mBIC-optimal. The obtained regime-switching dynamic appears to be persistent for all of the considered benchmarks. Particularly, the highest number of regime transitions allowed per regime is 21 in the case of DJI and the lowest number is 8 in the case of HSI. Using l_1 regularization led to identification of simpler models with respect to the number of regime parameters needed to describe the data. In all samples, except FTSE, we were able to eliminate irrelevant parameters, as reflected in the values of k^* of Table 4 representing the optimal number of the parameters in each regime. The corresponding regime densities with reduced parameters (shown in Figures 2 and 3) can be interpreted as curved exponential distributions. For the interpretation of the results it is important to note that the integration domain is finite and rescaled to the interval $[-1, 1]$ at the time of estimation. This approach allows us to resolve integrability issues otherwise present in fitting densities with maximum entropy. Apart from the MaxEnt distributions estimated by the method (in black), we fit the Gaussian densities to the data in each regime (in red). As shown in Table 2, the unconditional densities of all seven data sets are skewed with heavy tails. As seen from the comparison, the nonparametric TV-entropy densities provide a better fit for fat tails and asymmetry exhibited by the underlying densities, compared to the Gaussian distribution function.[5] This effect is most prominent in the regimes where the data points contributing to heavy mass at the tails are observed (for instance regime #3 of HSI in Figure 3).

As shown in Table 5, the memoryless TV-entropy model outperformed all of the finite-memory GARCH models across all seven benchmarks with respect to the

[5]The Gaussian density is fully described with two moments (e.g., the mean and the variance), and it corresponds to the MaxEnt distribution with the first two moment constraints ($k = 2$).

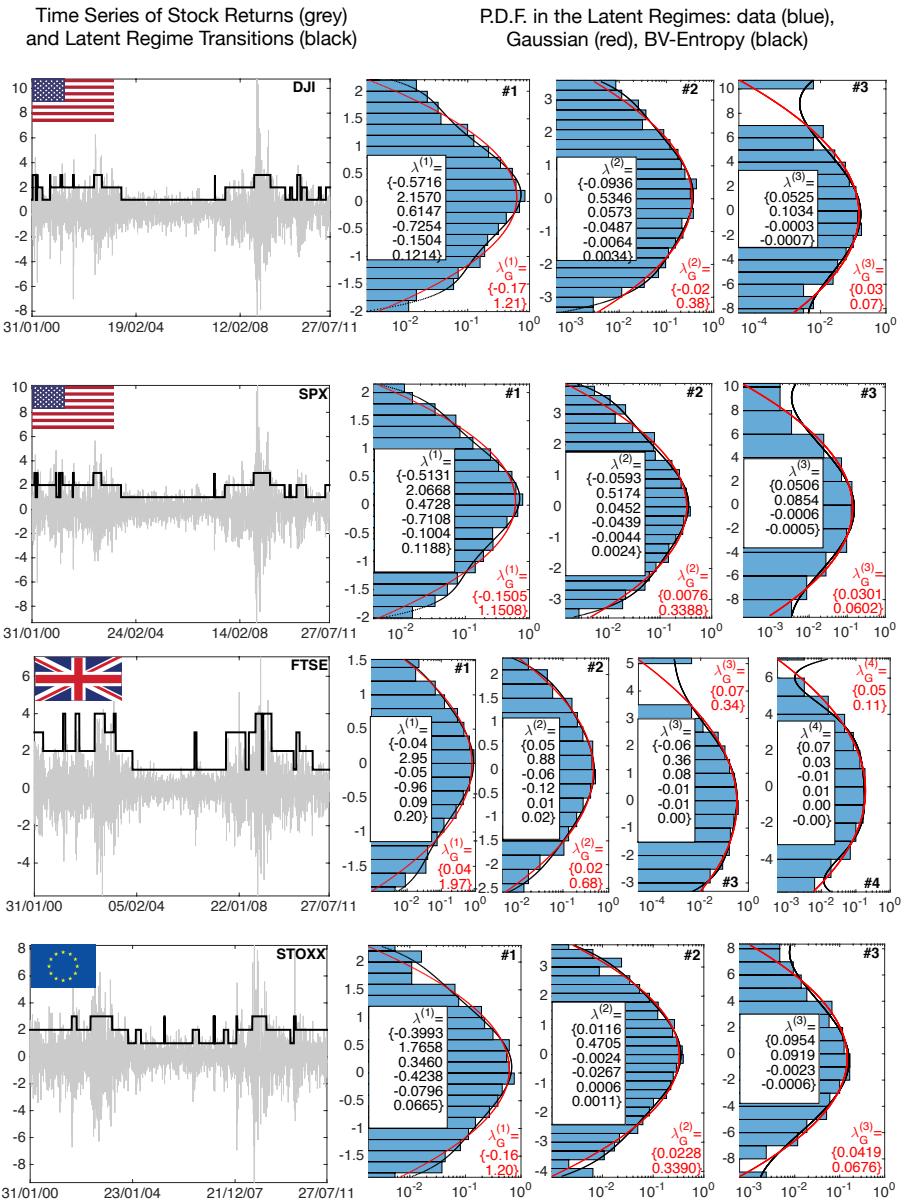

Figure 2. The optimal regime-switching paths obtained by the sparse multidimensional MaxEnt model (10)–(11) and the comparison of histograms of the corresponding regimes' data (blue bars) to the estimated MaxEnt densities (black) and fitted Gaussian densities (red) for DJI, SPX, FTSE, and STOXX index data.

log-likelihood and the BIC values. The highest log-likelihood value suggests the superior fit, while the lowest BIC value suggests the best balance between the fit and the complexity among all the considered models. As demonstrated by the posterior

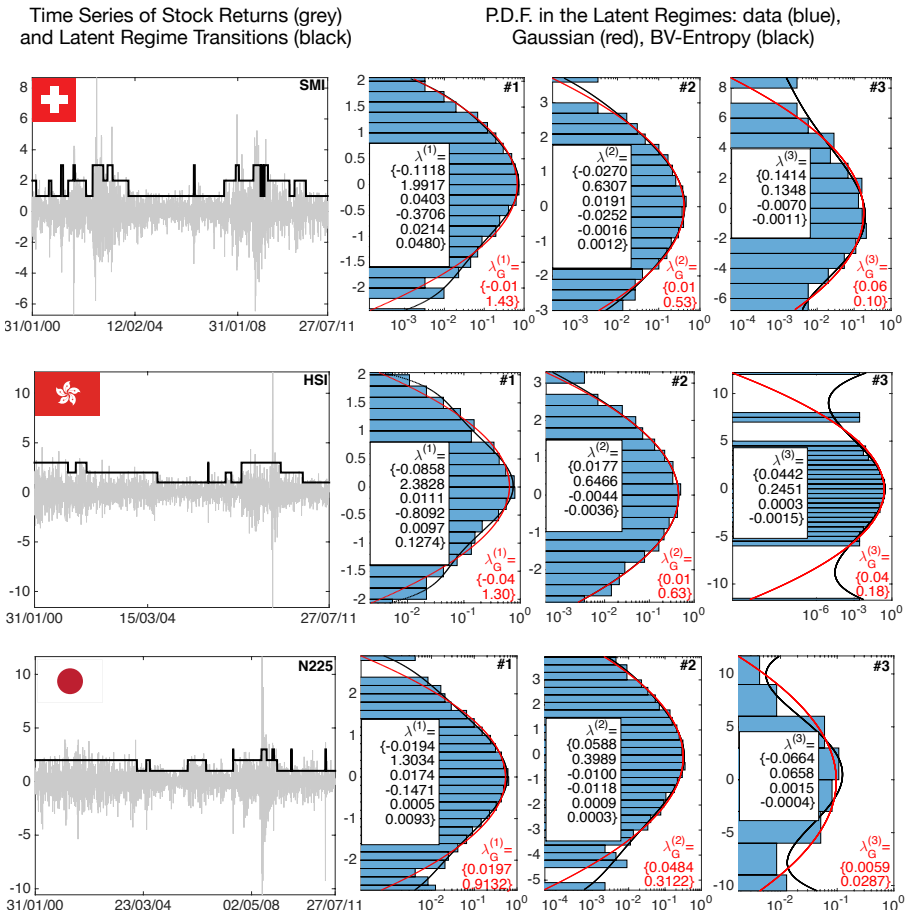

Figure 3. The optimal regime-switching paths obtained by the sparse multidimensional MaxEnt model (10)–(11) and the comparison of histograms of the corresponding regimes' data (blue bars) to the estimated MaxEnt densities (black) and fitted Gaussian densities (red) for SMI, HSI, and N225 index data.

probabilities inferred from Schwartz weights [6], the difference in BIC values is highly significant. These results indicate that the assumption of finite autoregressive memory imposed by the GARCH models is redundant for all of the considered time series data, once nonstationary regime switches are taken into account.

One of the central arguments for using the GARCH models is based on their ability to fit the autocorrelation function of $|x_t|^d$ as a function of lag times and exponents d [7]. As was shown in [7], there is little to no autocorrelation in daily returns. The highest autocorrelation is observed in the absolute returns and it is significant even at very large lags, suggesting a presence of autoregressive memory in the data. In the following we commence a simulation study where we draw

index	TV-entropy	GARCH	G-GJR	G-PML4	MS-GARCH
DJI	−3906.43	−4058.89	−4000.98	−4034.36	−4048.55
SPX	−4016.21	−4150.95	−4091.07	−4128.38	−4139.24
FTSE	−3465.43	−3580.84	−3552.85	−3563.42	−3577.93
STOXX	−4415.74	−4545.38	−4473.82	−4528.83	−4534.32
SMI	−3496.79	−3641.14	−3601.21	−3625.24	−3631.99
HSI	−3443.06	−3522.12	−3521.07	−3522.29	−3508.54
N225	−4003.36	−4100.27	−4080.22	−4067.87	−4094.51
DJI	7948.18	8141.66	8033.80	8108.52	8113.10
SPX	8167.73	8325.78	8213.99	8296.56	8294.47
FTSE	7129.99	7185.58	7137.57	7166.67	7171.86
STOXX	8966.99	9114.66	8979.52	9097.52	9084.64
SMI	7136.94	7306.16	7234.27	7290.30	7279.97
HSI	7004.08	7067.84	7073.59	7083.91	7047.08
N225	8141.50	8224.32	8192.15	8175.38	8205.02
DJI	(1.00)	(0.00)	(0.00)	(0.00)	(0.00)
SPX	(1.00)	(0.00)	(0.00)	(0.00)	(0.00)
FTSE	(0.98)	(0.00)	(0.02)	(0.00)	(0.00)
STOXX	(1.00)	(0.00)	(0.00)	(0.00)	(0.00)
SMI	(1.00)	(0.00)	(0.00)	(0.00)	(0.00)
HSI	(1.00)	(0.00)	(0.00)	(0.00)	(0.00)
N225	(1.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table 5. Log-likelihood (top), BIC (middle), and posterior probability (bottom) values obtained for in-sample analysis.

1000 samples from all previously obtained optimal models. We then compute the mean values of autocorrelation coefficients at each lag and compare them to the autocorrelation coefficients obtained on real data. Specifically, we analyze the autocorrelation function of $|x_t|^d$ as a function of d and a lag as shown in the left panels of Figure 4 for SMI and STOXX index data. The summary of results for all of the considered indices is shown in Table 6.

Consistently with [7], the highest serial correlation is observed around the value $d = 1$ for all of the considered benchmarks. As can be seen from Figure 4, even at the very large lags there is a significant serial correlation as the values of the coefficients lay outside the confidence intervals (light gray) constructed under the null hypothesis that the obtained series are completely random. Results from Figure 4 demonstrate that the simpler, conditionally memoryless TV-entropy models (green) provide the closest approximation of the sample autocorrelation function (gray), as compared to all of the considered GARCH models.

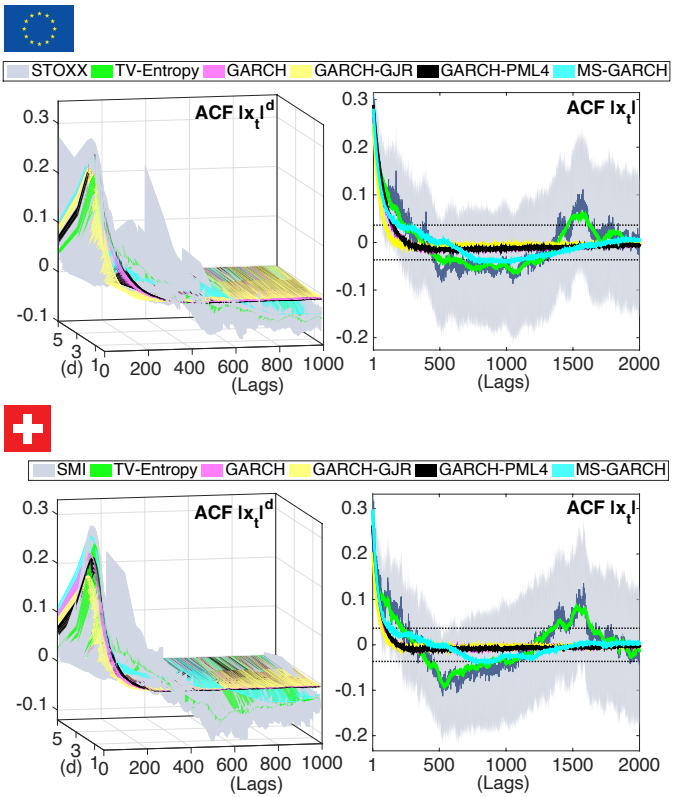

Figure 4. Autocorrelation function of the absolute returns as a function of lag, power d (left), and $d = 1$ (right) computed from the STOXX (top) and SMI (bottom) index data and the corresponding mean values inferred from the samples generated according to the optimal parameters of all considered models. The gray dashed lines are i.i.d. confidence intervals, and the gray shaded areas are confidence intervals under the assumption of the MA process.

Next, all of the obtained models will be used to produce volatility forecasts and, consequently, the one-day-ahead value-at-risk (VaR) forecasts. We compare the online VaR prediction quality for the considered methods and benchmarks. Once we observe the new data point, we confirm or correct the assigned regime value Γ using the optimal parameters obtained during the in-sample estimation. To compare the quality of the VaR forecasts, we commence the unconditional coverage test [20]. As shown in Table 7, there is no particular model that would be clearly preferred for all of the considered assets. However, since the results are not statistically distinguishable (the confidence intervals are overlapping), it can be concluded that the conditionally memoryless TV-entropy — relying on fewer tunable model parameters — is a good alternative to the traditional GARCH models in forecasting one-day-ahead VaR. Next, we analyze the relationships between the

index	TV-entropy	GARCH	GARCH-GJR	GARCH-PML4	MS-GARCH
DJI	✓ 0.00052	0.00120	0.00140	0.00088	0.00085
SPX	✓ 0.00037	0.00140	0.00180	0.00098	0.00073
FTSE	✓ 0.00024	0.00099	0.00210	0.00091	0.00065
STOXX	✓ 0.00034	0.00120	0.00170	0.00110	0.00073
SMI	✓ 0.00030	0.00190	0.00220	0.00180	0.00110
HSI	✓ 0.00036	0.00086	0.00093	0.00085	0.00062
N225	✓ 0.00041	0.00060	0.00063	0.00058	0.00084

Table 6. The mean squared error between sample and simulated autocorrelation coefficients of $|x_t|^d$ of the first 1000 lags for all considered models; ✓ denotes models with the best fit of the data autocorrelation.

index	TV-entropy	GARCH	G-GJR	G-PML4	MS-GARCH
DJI	0.044	0.042	0.043	0.041	0.043
SPX	0.045	0.045	0.047	0.042	0.046
FTSE	0.046	0.059	0.059	0.055	0.057
STOXX	0.054	0.052	0.052	0.048	0.052
SMI	0.048	0.049	0.051	0.047	0.049
HSI	0.056	0.054	0.051	0.054	0.015
N225	0.044	0.047	0.051	0.044	0.046
DJI	0.013	0.018	0.015	0.014	0.018
SPX	0.013	0.017	0.018	0.014	0.017
FTSE	0.015	0.021	0.023	0.017	0.021
STOXX	0.015	0.017	0.022	0.014	0.017
SMI	0.014	0.018	0.019	0.014	0.016
HSI	0.017	0.020	0.019	0.020	0.003
N225	0.013	0.016	0.017	0.012	0.019

Table 7. Unconditional coverage test results for 95% and 99% VaR. The expectation of VaR violations should be close to 5% and 1%, respectively.

latent volatility transitions inferred by TV-entropy (see the left panels of Figures 2 and 3). Identified MaxEnt regimes are characterized by different volatility levels. As a result, the regime transitions are jumps indicating an increase or a decrease of the volatility level. For every considered asset we construct two distinct binary variables describing such a behavior, where zeros indicate no jump at a current time instance, and ones stand for transition to the regime with only increased ("up") or only decreased ("down") volatility. We then perform a pairwise comparison of the obtained categorical time series (characterized by "up/up", "up/down", "down/up", and "down/down" directions) using Fisher's exact test [8] and analyze the resulting p-values to identify the statistically significant relations between the regime transitions.

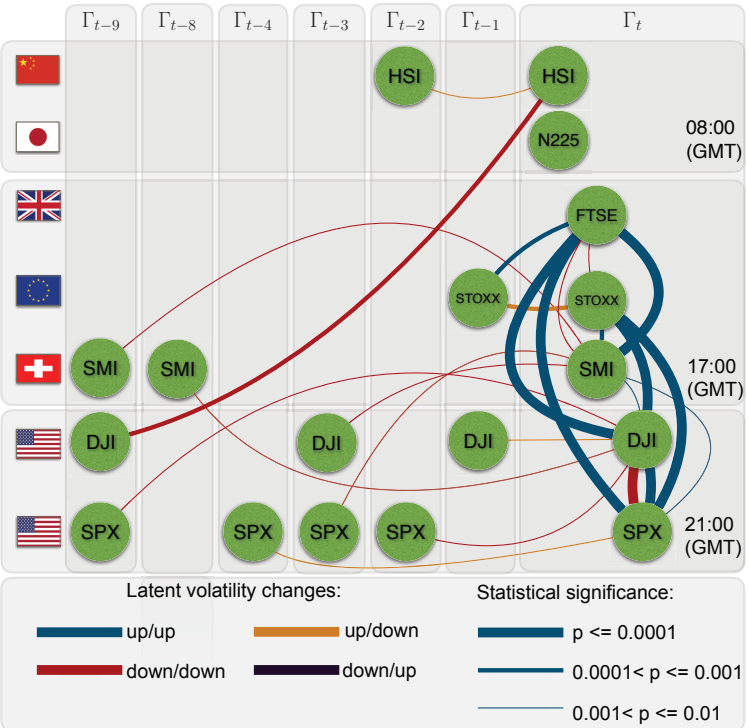

Figure 5. Latent structure relation graph, showing how the regime transitions are correlated between assets over time. The connection lines are scaled according to the p-value obtained by Fisher's exact test from the vales of latent regime variables Γ obtained with (10)–(11).

Shown in Figure 5 are only the most significant relationships ($p \leq 0.01$) among these categorical time series of latent regime transitions. The assets in the American and in the European regions appear to be heavily connected through their latent regime transitions. These connections are mostly governed by the volatility increase on the short time scale and volatility decrease on longer time scales. As for the considered Asian markets, the inferred latent influence of the American and European markets is present but it appears to be delayed in time and driven mostly by a decrease in the latent volatility levels.

Discussion

In this work we presented a sparse extension of the nonstationary MaxEnt methodology from one to multiple dimensions, aiming to identify the most qualitative (in terms of the log-likelihood) and the least complex (in terms of the information content and the required number of tunable parameters) representation for multidimensional time series data.

In an application to analysis of financial time series, we show that one of the important distinctions between the entropy-based and the common heteroscedastic approaches used in economics and finance is an assumption about the finite autoregressive memory in the underlying data-generating process. Traditionally, the observed data is assumed to be explicitly dependent on its past realizations. In the presented application of the MaxEnt framework, we do not impose additional assumptions about memory and do not include tunable parameters that describe it. The hypothesis that realizations are independent within the regimes has previously been explored in [5] and [28] in the context of parametric HMMs, where the authors showed that proposed models can reproduce the empirical properties of daily returns, especially in the case when conditional distributions are not normally distributed. However, these approaches also impose an explicit a priori memory assumption (Markovianity) on the level of the regime-switching process, and it remains unclear whether this assumption is necessary and/or sufficient for realistic financial data. The TV-entropy approach reveals that the volatility for all of the considered benchmarks is best described by the time-dependent persistent process, where persistence is identified through the adaptive regularization (11) of the regime-switching process and from the fact that within every regime the volatility remains stationary and i.i.d.

Starting with the intrinsically multivariate but ill posed MaxEnt formulation in (1)–(2), we derived that by introducing two (mild) Assumptions 1 and 2 its solution can be approximated from below with a well posed solution of the sparse regularized problem (10)–(11). Problem (10)–(11) is still multivariate since the hidden regime variables $\gamma_{t,i}^{(k)}$ change with the dimension i, time t, and regime k. As shown in Figure 5, they capture the latent multivariate relation structure over different dimensions and times.

In this manuscript we analyzed seven leading world market indices across America, Europe, and Asia. We found that the nonparametric TV-entropy approach outperforms all of the considered benchmark models for in-sample analysis in terms of the log-likelihood, simplicity (the number of free parameters), the information content (BIC and the posterior model probabilities), and the quality when describing the underlying autocorrelation function behavior. The out-of-sample study indicates that TV-entropy methodology is an effective alternative to the GARCH models for forecasting of the one-day-ahead VaR. The TV-entropy approach could closely reproduce serial correlation patterns found in data, especially at the large lags (unlike any of the GARCH models considered). This study indicates that nonstationary MS-GARCH (combining GARCH with the Markovian regime transition model) allows for a better description of the serial correlation than single-regime GARCH models, but it is not able to match the data as accurately as the TV-entropy model. Finally, Figure 5 illustrates how the regime transition processes inferred from the data allows us to identify the statistically significant temporal relations between

latent volatility level transitions across different markets. In particular, these findings indicate that negative news has a stronger short-term impact on the markets, as we observe that statistically most-significant latent connections are associated with the increase in volatility levels.

References

[1] N. Agmon, Y. Alhassid, and R. D. Levine, *An algorithm for finding the distribution of maximal entropy*, J. Comput. Phys. **30** (1979), no. 2, 250–258. Zbl

[2] L. Bauwens, A. Dufays, and J. V. K. Rombouts, *Marginal likelihood for Markov-switching and change-point GARCH models*, J. Econometrics **178** (2014), no. 2–3, 508–522. MR Zbl

[3] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, *A maximum entropy approach to natural language processing*, Comput. Linguist. **22** (1996), no. 1, 39–71.

[4] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, J. Econometrics **31** (1986), no. 3, 307–327. MR Zbl

[5] J. Bulla and I. Bulla, *Stylized facts of financial time series and hidden semi-Markov models*, Comput. Statist. Data Anal. **51** (2006), no. 4, 2192–2209. MR Zbl

[6] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd ed., Springer, 2002. MR Zbl

[7] Z. Ding, C. W. J. Granger, and R. F. Engle, *A long memory property of stock market returns and a new model*, J. Empir. Financ. **1** (1993), no. 1, 83–106.

[8] R. A. Fisher, *On the interpretation of χ^2 from contingency tables, and the calculation of p*, J. R. Stat. Soc. **85** (1922), no. 1, 87–94.

[9] S. Gerber and I. Horenko, *Improving clustering by imposing network information*, AAAS Sci. Adv. **1** (2015), no. 7, e1500163.

[10] L. R. Glosten, R. Jagannathan, and D. E. Runkle, *On the relation between the expected value and the volatility of the nominal excess return on stocks*, J. Financ. **48** (1993), no. 5, 1779–1801.

[11] A. Golan, G. Judge, and J. M. Perloff, *A maximum entropy approach to recovering information from multinomial response data*, J. Amer. Statist. Assoc. **91** (1996), no. 434, 841–853. MR Zbl

[12] L. P. Hansen, *Large sample properties of generalized method of moments estimators*, Econometrica **50** (1982), no. 4, 1029–1054. MR Zbl

[13] W. Härdle, *Applied nonparametric regression*, Econometric Society Monographs, no. 19, Cambridge University, 1990. MR Zbl

[14] G. Heber, A. Lunde, N. Shephard, and K. Sheppard, *Oxford-Man Institute's Realized Library*, Oxford University, 2009.

[15] A. Holly, A. Monfort, and M. Rockinger, *Fourth order pseudo maximum likelihood methods*, J. Econometrics **162** (2011), no. 2, 278–293. MR Zbl

[16] I. Horenko, *Finite element approach to clustering of multidimensional time series*, SIAM J. Sci. Comput. **32** (2010), no. 1, 62–83. MR Zbl

[17] ———, *On the identification of nonstationary factor models and their application to atmospheric data analysis*, J. Atmos. Sci. **67** (2010), no. 5, 1559–1574.

[18] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. (2) **106** (1957), 620–630. MR Zbl

[19] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, IJCAI '95: proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, Morgan Kaufmann, San Francisco, 1995, pp. 1137–1143.

[20] P. H. Kupiec, *Techniques for verifying the accuracy of risk measurement models*, J. Deriv. **3** (1995), no. 2, 73–84.

[21] C. F. Manski and D. McFadden (eds.), *Structural analysis of discrete data with econometric applications*, MIT, Cambridge, MA, 1981. MR Zbl

[22] G. Marchenko, P. Gagliardini, and I. Horenko, *Towards a computationally tractable maximum entropy principle for nonstationary financial time series*, SIAM J. Financial Math. **9** (2018), no. 4, 1249–1285. MR Zbl

[23] L. R. Mead and N. Papanicolaou, *Maximum entropy in the problem of moments*, J. Math. Phys. **25** (1984), no. 8, 2404–2417. MR

[24] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Jr., *Maximum entropy models for antibody diversity*, P. Natl. Acad. Sci. USA **107** (2010), no. 12, 5405–5410.

[25] K. Nigam, J. Lafferty, and A. McCallum, *Using maximum entropy for text classification*, IJCAI '99: workshop on machine learning for information filtering, 1999, pp. 61–67.

[26] S. J. Phillips, R. P. Anderson, and R. E. Schapire, *Maximum entropy modeling of species geographic distributions*, Ecol. Model. **190** (2006), no. 3–4, 231–259.

[27] L. Pospíšil, P. Gagliardini, W. Sawyer, and I. Horenko, *On a scalable nonparametric denoising of time series signals*, Commun. Appl. Math. Comput. Sci. **13** (2018), no. 1, 107–138. MR Zbl

[28] T. Rydén, T. Teräsvirta, and S. Åsbrink, *Stylized facts of daily return series and the hidden Markov model*, J. Appl. Economet. **13** (1998), no. 3, 217–244.

[29] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288. MR Zbl

[30] X. Wu, *Calculation of maximum entropy densities with application to income distribution*, J. Econometrics **115** (2003), no. 2, 347–354. MR Zbl

[31] A. Zellner and R. A. Highfield, *Calculation of maximum entropy distributions and approximation of marginal posterior distributions*, J. Econometrics **37** (1988), no. 2, 195–209. MR

Received October 8, 2019. Revised May 6, 2020.

ILLIA HORENKO: horenkoi@usi.ch
Institute of Computational Science, Facoltà di scienze informatiche, Università della Svizzera italiana, Lugano, Switzerland

GANNA MARCHENKO: ganna.marchenko@usi.ch
Institute of Computational Science, Facoltà di scienze informatiche, Università della Svizzera italiana, Lugano, Switzerland

PATRICK GAGLIARDINI: patrick.gagliardini@usi.ch
Facoltà di scienze economiche Università della Svizzera italiana, Lugano, Switzerland

and

Swiss Finance Institute, Zürich, Switzerland

Communications in Applied Mathematics and Computational Science

msp.org/camcos

See inside back cover or msp.org/camcos for submission instructions.

The subscription price for 2020 is US $110/year for the electronic version, and $165/year (+$15, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscriber address should be sent to MSP.

Communications in Applied Mathematics and Computational Science (ISSN 2157-5452 electronic, 1559-3940 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

CAMCoS peer review and production are managed by EditFLOW® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

Communications in Applied Mathematics and Computational Science

vol. 15 no. 2 2020

Simulating single-coil MRI from the responses of multiple coils 115
Mark Tygert and Jure Zbontar

On a computationally scalable sparse formulation of the multidimensional and nonstationary maximum entropy principle 129
Illia Horenko, Ganna Marchenko and Patrick Gagliardini

Fast multigrid solution of high-order accurate multiphase Stokes problems 147
Robert I. Saye