SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

September 2018, Vol. 61 098106:1–098106:3
https://doi.org/10.1007/s11432-017-9335-5

Predicting compositional time series via autoregressive Dirichlet estimation

Ganbin ZHOU[1,2], Ping LUO[1,2] & Qing HE[1,2*]

[1] *Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China;*
[2] *University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received 26 July 2017/Revised 29 November 2017/Accepted 8 January 2018/Published online 18 May 2018

Citation Zhou G B, Luo P, He Q. Predicting compositional time series via autoregressive Dirichlet estimation. Sci China Inf Sci, 2018, 61(9): 098106, https://doi.org/10.1007/s11432-017-9335-5

In recent years, compositional time series (CTS) prediction has become a widely applied data analysis method for modeling tactile sequence data [1], hydrological time series data using a four-stage algorithm (denoising, decomposition, components prediction and ensemble) [2], and daily and monthly extreme temperature data [3, 4]. In general, CTS data consists of a sequence of K-dimensional vectors, each of which contains proportion information for a specific time point [5]. Furthermore, it is required that $\sum_{j=1}^{K}\theta_{i,j} = 1$. Thus, the aim of this study is to predict $\boldsymbol{\theta}_t$ based on $\{\boldsymbol{\theta}_i\}_{i=1}^{t-1}$, which is a common problem in CTS prediction.

We chose the Chinese college entrance examination (CCEE) as a real-world application for our proposed method. Thus, we first revisited earlier studies on CTS prediction, and found that all previous methods can be grouped into two classes: those with data transformation and those without. The first of these classes canceled the constraint that the sum of all the components in a vector must equal 1, transformed the data to unconstrained vectors, and then adopted traditional time series analysis methods for prediction [5]. However, this data transformation tended to alter the nature of the original compositional data. The second class of methods handled CTS data directly. For example, methods such as CDES [6] and DRHT [7] independently model the trend of each dimension of CTS data via a smoothing or regression method; however, when the CTS data are volatile, performance was unsatisfactory. In addition, methods such as VARMA [5] focused on modeling the correlations between the different dimensions of CTS data. However, in real-world applications, the CTS data usually contain many dimensions and few data points.

Observations on different CTS data types. After careful study, we find two types of CTS data produced using different data generation mechanisms. The first of these methods randomly generates CTS data using latent Dirichlet distribution. One example of this CTS data type is the distribution of knowledge test points in the CCEE. Each year, exam setters craft questions according to the most recent official exam outline published by the government. Knowledge points that are considered “hot” one year may not be considered “hot” the following year, indicating that internal regularities or trends in these data may not exist; however, such regularities and trends may exist in other types of CTS data (such as economic structure, population, and social classes). The major challenge is that internal regularities trends vary for different CTS data types, which leads to difficulties in accurately representing the data of these two situations within a single model. Hence, we first propose two models for these CTS data types: Dirichlet estimation and probabilistic autoregres-

[*] Corresponding author (email: heqing@ict.ac.cn)

sion. Furthermore, we argue that good prediction performance for both types of CTS data can be achieved by a amalgamation between these two models, and thus propose a joint model.

Predicting CTS via Dirichlet estimation. We will now focus on our proposed method of predicting $\boldsymbol{\theta}_t$ for the CTS sequence $\{\boldsymbol{\theta}_i\}_{i=1}^{t-1}$. The first model is based on Dirichlet estimation and is suitable for CTS data that have been generated randomly using latent Dirichlet distribution.

Note that the support for a K-dimensional Dirichlet distribution is $S^{K-1} = \{\boldsymbol{\theta} \in [0, +\infty)^K \, | \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{I}_K = 1\}$, where $\boldsymbol{I}_K = (1, 1, \ldots, 1)^{\mathrm{T}}$. S^{K-1} is a $(K-1)$-dimensional simplex. Dirichlet distribution is an exponential family distribution method, meaning that it has a conjugate prior. Let $D(\boldsymbol{\theta}; \boldsymbol{\beta})$ be the probability distribution function (PDF) of the Dirichlet distribution with the parameters $\boldsymbol{\beta}$, and $\pi(\boldsymbol{\beta}; \boldsymbol{d}, v)$ be the exponential family prior distribution of $\boldsymbol{\beta}$ with the parameters $(\boldsymbol{d}, v)$. As Dirichlet distribution can directly model compositional data, we assume that the observations in a recent short-term window, namely $\{\boldsymbol{\theta}_i\}_{i=t-s}^{t}$, are generated from a background Dirichlet distribution with the parameters $\boldsymbol{\beta}_t$. In other words, we have $\{\boldsymbol{\theta}_i\}_{i=t-s}^{t} \sim D(\boldsymbol{\theta}; \boldsymbol{\beta}_t)$ where $D(\boldsymbol{\theta}; \boldsymbol{\beta}) = \frac{1}{B(\boldsymbol{\beta})} \exp\{(\boldsymbol{\beta} - 1)^{\mathrm{T}} \ln \boldsymbol{\theta}\}$ and $s \in \mathbb{N}_+$ is the size of this short-term window.

Furthermore, we denote the prior of $\boldsymbol{\beta}_t$ as $\pi(\boldsymbol{d}_t, v_t)$. Following this, according to the properties of conjugate prior, we derived the posterior PDF of $\boldsymbol{\beta}_t$ via the likelihood that, $P(\boldsymbol{\beta}_t | \{\boldsymbol{\theta}_i\}_{i=t-s}^{t-1}, \boldsymbol{d}_t, v_t) = \pi(\boldsymbol{\beta}_t; \boldsymbol{d}_t + \sum_{i=t-s}^{t-1} \ln \boldsymbol{\theta}_i, v_t + s)$ where $\pi(\boldsymbol{\beta}; \boldsymbol{d}, v) = \frac{r(\boldsymbol{d}, v)}{B^v(\boldsymbol{\beta})} \exp\{(\boldsymbol{\beta} - 1)^{\mathrm{T}} \boldsymbol{d}\}$.

However, the hyperparameters $\boldsymbol{d}_t$ and v_t were still unknown. Thus, we assumed that $\boldsymbol{d}_t$ and v_t were determined via observations in a long-term time window, namely $\{\boldsymbol{\theta}_i\}_{i=t-l}^{t-1}$, where $l \in \mathbb{N}_+$ was the size of this long-term window and $l > s$. Specifically, $\boldsymbol{\beta}_t \sim \pi(\boldsymbol{\beta}_t; \boldsymbol{d}_t, v_t) = \pi(\boldsymbol{\beta}_t; \sum_{i=t-l}^{t-1} \ln \boldsymbol{\theta}_i, l)$.

This result was derived as follows: there was a non-informative prior $\boldsymbol{\beta}_0$ [8] for the observations before time point $(t-l)$. Following this, the posterior of $\boldsymbol{\beta}_0$ on observations $\{\boldsymbol{\theta}_i\}_{i=t-l}^{t-1}$ was $P(\boldsymbol{\beta}_0 | \{\boldsymbol{\theta}_i\}_{i=t-l}^{t-1}) = \pi(\boldsymbol{\beta}_0; \sum_{i=t-l}^{t-1} \ln \boldsymbol{\theta}_i, l)$. Here, we assumed the posterior of $(\boldsymbol{\beta}_0; \{\boldsymbol{\theta}_i\}_{i=t-l}^{t-1})$ to be the prior of $\boldsymbol{\beta}_t$. Thus, $\boldsymbol{\beta}_t \sim \pi(\boldsymbol{\beta}_t; \boldsymbol{d}_t, v_t) = \pi(\boldsymbol{\beta}_t; \sum_{i=t-l}^{t-1} \ln \boldsymbol{\theta}_i, l)$, and we have $\boldsymbol{d}_t = \sum_{i=t-l}^{t-1} \ln \boldsymbol{\theta}_i$ and $v_t = l$.

The objective Dirichlet estimation function is as follows: $O_D(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t) = P(\boldsymbol{\theta}_t | \{\boldsymbol{\theta}_i\}_{i=t-l}^{t-1}, l, s) = D(\boldsymbol{\theta}_t; \boldsymbol{\beta}_t) \, \pi(\boldsymbol{\beta}_t; \sum_{i=t-l}^{t-1} \ln \boldsymbol{\theta}_i + \sum_{i=t-s}^{t-1} \ln \boldsymbol{\theta}_i, l + s)$.

Here, we maximized the above objective function to the range of $\boldsymbol{\theta}_t \in S^{K-1}$ and $\boldsymbol{\beta}_t \in \mathbb{R}_+^K$ [9].

Predicting CTS via probabilistic autoregression. We now move to the second proposed model, in which CTS data are generated by internal regularities or trends. First, we adapted an autoregressive model for CTS data. Following this, we converted this autoregressive model into a probabilistic model so that it could be easily combined with the first proposed model. In autoregressive model, $\boldsymbol{\theta}_t$ was estimated by using the linear weighted sum of $\{\boldsymbol{\theta}_i\}_{i=t-s}^{t-1}$. Specifically, let $K \times s$ block matrix $\boldsymbol{\Theta}_i = [\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i-1}, \ldots, \boldsymbol{\theta}_{i-s+1}]$, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_s)^{\mathrm{T}}$. Thus,

$$\begin{cases} \boldsymbol{\theta}_t = \displaystyle\sum_{i=1}^{s} \lambda_i \boldsymbol{\theta}_{t-i} + \boldsymbol{\epsilon}_t = \boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda} + \boldsymbol{\epsilon}_t, \quad t \geqslant s, \\ \boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda} \in S^{K-1}, \end{cases} \tag{1}$$

where $\boldsymbol{\lambda}$ is the autoregressive coefficient vector, and vector $\boldsymbol{\epsilon}_t$ is the error item.

Following this, the coefficient vector $\boldsymbol{\lambda}$ was estimated by minimizing the error item. This traditionally assumes that $\boldsymbol{\epsilon}_t$ follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$; however, it is difficult to determine the covariance matrix $\boldsymbol{\Sigma}$ when the sequence is short. Here, we assumed that $\|\boldsymbol{\epsilon}_t\|_2 \sim \mathcal{N}(0, \sigma^2)$ and it lay within the interval $[0, \sqrt{2}]$. In addition, the L^2-norm of $\boldsymbol{\epsilon}_t$ was used for convenience in the following computing, and $\|\boldsymbol{\epsilon}_t\|_2$ conditional on $[0, \sqrt{2}]$, had a truncated normal distribution.

Here, the upper bound on this truncated normal distribution was $\sqrt{2}$, which was derived as follows:

$$\begin{aligned} \|\boldsymbol{\epsilon}_t\|_2^2 &= \|\boldsymbol{\theta}_t - \boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda}\|_2^2 \\ &= \boldsymbol{\theta}_t^2 - 2\boldsymbol{\theta}_t^{\mathrm{T}} (\boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda}) + (\boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda})^2 \\ &\leqslant \boldsymbol{\theta}_t^2 + (\boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda})^2 \leqslant 1 + 1 = 2. \end{aligned} \tag{2}$$

This equality holds only if $\|\boldsymbol{\theta}_t\|_2 = 1$, $\|\boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda}\|_2 = 1$, and $\boldsymbol{\theta}_t$ are orthogonal to $\boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda}$. Hence, the PDF of $\|\boldsymbol{\epsilon}_t\|_2$ was as follows:

$$P(\|\boldsymbol{\epsilon}_t\|_2; \sigma^2) = \frac{\phi(\frac{\|\boldsymbol{\epsilon}_t\|_2}{\sigma})}{\Phi(\frac{\sqrt{2}}{\sigma}) - \Phi(\frac{0}{\sigma})} \propto \exp\left\{\frac{\|\boldsymbol{\epsilon}_t\|_2^2}{-2\sigma^2}\right\}, \tag{3}$$

where $\phi(\cdot)$ is the PDF of standard Gaussian distribution, $\Phi(\cdot)$ is its cumulative distribution function, and $\sigma \in \mathbb{R}_+$ is a chosen scalar. Given that $\boldsymbol{\epsilon}_t = \boldsymbol{\theta}_t - \boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda}$, $\boldsymbol{\epsilon}_t$ is the function of $\boldsymbol{\lambda}$. Thus, we have $P(\|\boldsymbol{\epsilon}_t\|_2; \sigma^2, \boldsymbol{\lambda}) \propto \exp\{\frac{(\boldsymbol{\theta}_t - \boldsymbol{\Theta}_{t-1} \boldsymbol{\lambda})^2}{-2\sigma^2}\}$.

Finally, given $\{\boldsymbol{\theta}_i\}_{i=t-m-s}^{t-1}$, we can estimate both $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}_t$ by maximizing the following likelihood:

$$O_A(\boldsymbol{\theta}_t, \boldsymbol{\lambda}) = \prod_{i=t-m}^{t} P(\|\boldsymbol{\epsilon}_i\|_2; \sigma^2, \boldsymbol{\lambda})$$

$$\propto \exp\left\{\frac{(\boldsymbol{\theta}_t - \boldsymbol{\Theta}_{t-1}\boldsymbol{\lambda})^2}{-2\sigma^2}\right\} \cdot \prod_{i=t-m}^{t-1} P(\|\boldsymbol{\epsilon}_i\|_2; \sigma^2, \boldsymbol{\lambda}), \qquad (4)$$

where m is a chosen scalar. Here, as the likelihood increases, the error terms $\{\|\boldsymbol{\epsilon}_i\|_2\}_{i=t-m}^{t}$ decrease.

Amalgamating Dirichlet estimation and probabilistic autoregression. We have proposed two models for two CTS data types. For data that are independently generated by latent Dirichlet distribution, the method of Dirichlet estimation was effective. For the data that change under internal regularities or trends, the method of probabilistic autoregression was effective.

In order to improve the robustness and efficacy of our proposed model for unknown CTS data types, we propose the following joint model, which was created by multiplying the two objective functions. After some derivation, we have

$$\begin{aligned} &O_D(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t) \times O_A(\boldsymbol{\theta}_t, \boldsymbol{\lambda}) \\ &\propto \exp\left\{(\boldsymbol{\beta}_t - 1)^{\mathrm{T}}\left[\ln(\boldsymbol{\theta}_t) + \sum_{i=t-l}^{t-1}\ln(\boldsymbol{\theta}_i)\right.\right. \\ &\left.\left.+ \sum_{i=t-s}^{t-1}\ln(\boldsymbol{\theta}_i)\right] - \frac{1}{2\sigma^2}\sum_{i=t-m}^{t}(\boldsymbol{\theta}_i - \boldsymbol{\Theta}_{i-1}\boldsymbol{\lambda})^2\right\} \\ &\cdot \frac{1}{B^{l+s+1}(\boldsymbol{\beta}_t)}. \end{aligned} \qquad (5)$$

The model structure is shown in Figure 1. Eq. (5) is the final objective function for this joint model and is called the autoregressive Dirichlet estimation (ADE).

$$(\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\lambda}}) = \underset{(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\lambda})}{\operatorname{argmax}} f(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \boldsymbol{\lambda}). \qquad (6)$$

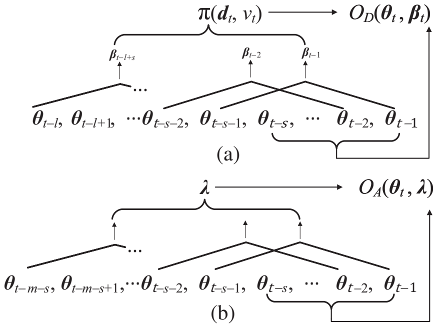
π(d_t, v_t)
O_D(θ_t, β_t)
β_{t−l+s}
β_{t−2}
β_{t−1}
θ_{t−l}, θ_{t−l+1}, ···θ_{t−s−2}, θ_{t−s−1}, θ_{t−s}, ··· θ_{t−2}, θ_{t−1}
(a)
λ
O_A(θ_t, λ)
θ_{t−m−s}, θ_{t−m−s+1}, ···θ_{t−s−2}, θ_{t−s−1}, θ_{t−s}, ··· θ_{t−2}, θ_{t−1}
(b)

Figure 1 The model structure. (a) The autoregressive Gaussian estimation for CTS; (b) the probabilistic autoregression for CTS.

In (5), the parameter σ controls the importance tradeoff between the original two objectives. The autoregression weight increases as $\frac{1}{2\sigma^2}$ increases, and vice versa; therefore, we must set $\frac{1}{2\sigma^2}$ higher for the second CTS data type and lower for the first CTS data type. The experimental results demonstrate that we can improve performance of the proposed model if σ is set according to this rule.

Experiments. We compared our proposed ADE model with three state-of-the-art baseline methods, and prepared two datasets for experiments: CCEE and world development indicators. The experimental details are presented in Appendix A.

Conclusion. We first proposed two methods: Dirichlet estimation and probabilistic autoregression. Following this, we proposed a joint ADE model, which is a amalgamation between our original two methods. Our experiments demonstrated the performance of the joint ADE model for different CTS data types.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2017YFB1002104), National Natural Science Foundation of China (Grant Nos. 617733 61, 91546122, 61573335, 61602438, 61473273, 61473 274), and Guangdong Provincial Science and Technology Plan Projects (Grant No. 2015 B010109005).

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Liu H, Guo D, Sun F C. Object recognition using tactile measurements: kernel sparse coding methods. IEEE Trans Instrum Meas, 2016, 65: 656–665

2 Di C L, Yang X H, Wang X C. A four-stage hybrid model for hydrological time series forecasting. PLoS ONE, 2014, 9: e104663

3 Zhang J, Yang X H, Chen X J. Wavelet network model based on multiple criteria decision making for forecasting temperature time series. Math Prob Eng, 2015, 2015: 1–4

4 Zhang J, Yang X H, Li Y Q. A refined rank set pair analysis model based on wavelet analysis for predicting temperature series. Int J Numer Method Heat Fluid Flow, 2015, 25: 974–985

5 Mills T C. Forecasting compositional time series. Qual Quant, 2010, 44: 673–690

6 Raharjo H, Xie M, Brombacher A C. On modeling dynamic priorities in the analytic hierarchy process using compositional data analysis. Eur J Oper Res, 2009, 194: 834–846

7 Wang H W, Liu Q, Mok H M K, et al. A hyperspherical transformation forecasting model for compositional data. Eur J Oper Res, 2007, 179: 459–468

8 Gelman A. Prior distributions for variance parameters in hierarchical models. Econometrics, 2006, 1: 515–534

9 Thomas M. Estimating a Dirichlet Distribution. Technical Report MIT, 2000