Bayesian non-parametric approaches for dependent processes

Kaspar Märtens Xenia Miscouridou Paul Vanetti

March 10, 2016

Abstract

Dependent non-parametric random processes extend the notion of random measures from distributions over individual measures to distributions over collections of measures. This is done by considering a family of random probability measures, indexed by some covariate which will typically correspond to space or time. Many constructions of dependent processes are based on extensions of the Dirichlet process. We focus on two such extensions: the probit stick-breaking process, and the spatial normalized gamma process.

1 Introduction

Non-parametric priors are popular in statistics and machine learning for their capacity to represent large classes of models. A famous example is the Dirichlet process (DP) which is commonly used as a prior over clusters. In this context, the model does not require the prior specification of the number of clusters, but instead results in a posterior which includes a random number of clusters.

The DP exhibits *exchangeability*, which is desirable in settings where the ordering of data is arbitrary. Unfortunately, extending the DP while maintaining this property is difficult, the Pitman-Yor process [5] being the main example. As such, many extensions have been proposed which break this exchangeability property, see [2] for a comprehensive survey.

These non-exchangeable extensions are appropriate in some settings where the data are not believed to be exchangeable; we shall focus primarily on the common *time-series* setting, in which case the data are associated with specific times.

Dependent random processes offer the flexibility to depend on covariates and hence allow processes share parameters between them. The goal of dependent non-parametric processes is to generalize non-parametric priors over measures, partitions and sequences to priors over collections of such random structures. Typically, the closer the covariates of two processes are (in covariate space), the greater the degree of dependency between the processes and the greater the similarity of their behaviour.

2 Non-parametric models for exchangeable data

In Bayesian non-parametric models, the parameters $\theta \in \Theta$ are modeled as infinite-dimensional, allowing the model complexity (i.e. the finite subset of θ used for modeling observations) to grow with the sample size. Usually, it is assumed that observations are exchangeable, and it is of interest to construct distributions over discrete probability measures on Θ. One commonly-used prior is the Dirichlet process.

2.1 Dirichlet process

Denote the Dirichlet process (DP) as $\mathrm{DP}(\alpha_0, G_0)$, with concentration parameter $\alpha_0 > 0$ and base measure G_0 on Θ. The DP is defined as the distribution of the probability measure G such that for any finite measurable partition $(A_1, \ldots, A_r)$ of Θ the vector $(G(A_1), \ldots, G(A_r))$ is distributed according to the Dirichlet distribution

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)).$$

It turns out that $G \sim \mathrm{DP}(\alpha_0, G_0)$ is an atomic measure $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$, where the atoms' locations θ_k are distributed i.i.d. according to G_0, and their weights π_k can be obtained via a stick-breaking process, i.e. $\pi_k = V_k \prod_{i<k}(1 - V_i)$, where $V_k \overset{iid}{\sim} \mathrm{Beta}(1, \alpha_0)$.

2.2 Completely Random Measures and Normalized Random Measures

The DP is an example of a *normalized random measure* (NRM). In general, we can divide the class of random measures in two subclasses: *completely random measures* (CRM's) and NRM's. Following the notation above, a CRM is a distribution over measures on some measurable space such that for disjoint measurable sets $(A_1, \ldots, A_r)$, the random variables $G(A_1), \ldots, G(A_r)$ are independent. Moreover, a CRM has a Poisson process representation and can thus be realized by simulating a non-homogeneous Poisson process with the appropriate rate measure. An NRM is a distribution over probability measures and can be obtained by normalizing the output of a CRM.

2.3 Exchangeability

An exchangeable sequence is one whose joint distribution is invariant to permutations and every such sequence has a representation theorem provided by de Finetti (see e.g. [3]) which states that any infinitely exchangeable sequence can be written as a mixture of i.i.d. samples. Analogous representation theorems exist for exchangeable partitions and matrices. Under the assumption of exchangeable observations, non-parametric priors such as the Dirichlet process (DP), the Chinese restaurant process (CRP), the Pitman-Yor process, or the Indian buffet process (IBP) are appropriate.

One such process which deviates from the classical exchangeable setting is the *hierarchical Dirichlet process* (HDP). It is suitable in settings where the observations are organized into groups and assumed to be exchangeable both within each group and across groups. The models associated with each group are linked together and the HDP is used as a prior in this grouped mixture model setting. HDP's however are not the primary focus of the project but other sorts of dependencies on random measures.

3 Non-parametric models for dependent data

Often, there is a need to model data containing spatial or time dependencies, but the models considered in the previous section assume exchangeability. To incorporate dependencies, one possible approach is to replace a single non-parametric process with a collection of processes. That is, for a fixed covariate x, we replace G with a collection $\{G^{(x)} : x \in \mathcal{X}\}$. For instance, for continuous time we take $\mathcal{X} = \mathbb{R}^+$, and for spatial data $\mathcal{X} = \mathbb{R}^2$. One can take different approaches

to introduce dependency between these random measures, and there is no single straight-forward way to do this. MacEachern originally proposed a set of criteria for a distribution over this collection of measures [4]; many models have since been proposed for this purpose.

In principle, for each $G^{(x)}$, its locations θ_k and weights π_k may depend on x, i.e.

$$G^{(x)} = \sum_{k=1}^{\infty} \pi_k^{(x)} \delta_{\theta_k^{(x)}}$$

Therefore, two approaches that arise naturally for introducing dependency between these measures are:

- dependence on atom location, i.e. the weights $\pi_k^{(x)} = \pi_k$ for all x are shared
- dependence on atom weights, i.e. their locations $\theta_k^{(x)} = \theta_k$ for all x are shared

For example, the latter may be achieved via a suitable stick-breaking scheme (see Section 3.2).

Instead of directly modifying the atom locations or their weights, a set of dependent random measures may be constructed starting from a CRM or from its equivalent Poisson process representation, and via normalization the NRM is obtained. This will be discussed further in Section 3.1.

3.1 Dependence via Completely Random Measures

In order to construct any set of dependent random measures it is natural to start from a CRM or equivalently from its Poisson process representation. Then, any operation in the Poisson process will yield a CRM and vice-versa. Allowing the operation to depend on a covariate we then define a dependent CRM that varies with that covariate and by normalization we recover a NRM. Dependent Dirichlet Processes (DDP) define a measure on collections of measures which are dependent sets of random measures. The dependent measures are marginally DP's which is an important property. In general, DDP's have a number of desirable properties as described in [4]. The four properties are:

1. the support of the distribution on the measures $D_{s_1}, ..., D_{s_k}$ should be large,
2. when the distribution is used in a Bayesian hierarchical model it should be amenable to updating,
3. every measure D_s should marginally follow a known distribution (DP for the case of DDP's), and
4. measures drawn from the distribution should converge in the sense that $D_s \to D_{s_0}$ as $s \to s_0$.

3.1.1 Spatial Normalized Gamma Processes

One way to construct dependent DP's is by marginalizing and normalizing gamma processes over an extended space, as described in the spatial normalized gamma process (SNΓP) of [6]. The idea is to define a gamma process over an extended space; by taking any subset of this space and normalizing, a DP is obtained. The degree of dependence among these DP's is induced by the amount of overlap among the regions.

In order to connect these DP's to our model, we take some covariate t_j and associate a region of the space with this covariate; call this region Y_{t_j}. The DP associated with observation t_j, which we denote D_{t_j}, is found by normalizing the gamma process over the space Y_{t_j}.

The covariate need not be discrete, but a finite number of observations will induce only a finite number of subsets of the extended space (i.e. the Y_{t_j}) which much be considered. As such, we may divide the space into a partition from which these subsets may be constructed: let $\mathcal{R}$ denote the smallest collection of disjoint regions in the extended space so that each Y_{t_j} is a union of regions in $\mathcal{R}$. Let R_j be those regions which comprise Y_{t_j}, i.e. $Y_{t_j} = \cup_{R \in R_j} R$. This set can be interpreted as a mixture of independent DP's defined on the disjoint regions R_j.

Using the partitioning scheme described above, the construction for a SNΓP can be given in the following form; here α_R is a measure on Θ, which may be obtained from the measure on the extended space by marginalizing out the variables associated with the construction of the partitions.

$$\begin{aligned} g_R &\sim \text{Gamma}(\alpha_R(\Theta)) \\ D_R &\sim DP(\alpha_R) \\ D_{t_j} &= \sum_{R \in R_j} \frac{g_R}{\sum_{R' \in R_j} g_{R'}} D_R \end{aligned}$$

In Figure 1, we demonstrate a prior sample from a construction described in [6]. This construction, designed for time-series applications, constructs a partition spanning every interval formed by two time points. That is, there exist regions R_{t_a,t_b} for every interval between observed times $t_a \leq t_b$ and the regions for a given observation are those where the interval contains t_j: $R_j = \cup\{R_{t_a,t_b} : t_a \leq t_j \leq t_b\}$. In this example, we have only $t \in \{1, 2\}$, so there are only three possible intervals. Figure 2 shows the equivalent diagram for three time points; note the $O(n^2)$ growth of the number of regions.

We have introduced one example of an SNΓP - it is important to note that in general the SNΓP leaves a great deal of freedom in choosing the dependency scheme, and we could pick more or less complicated models depending on our modeling goals and computation considerations.

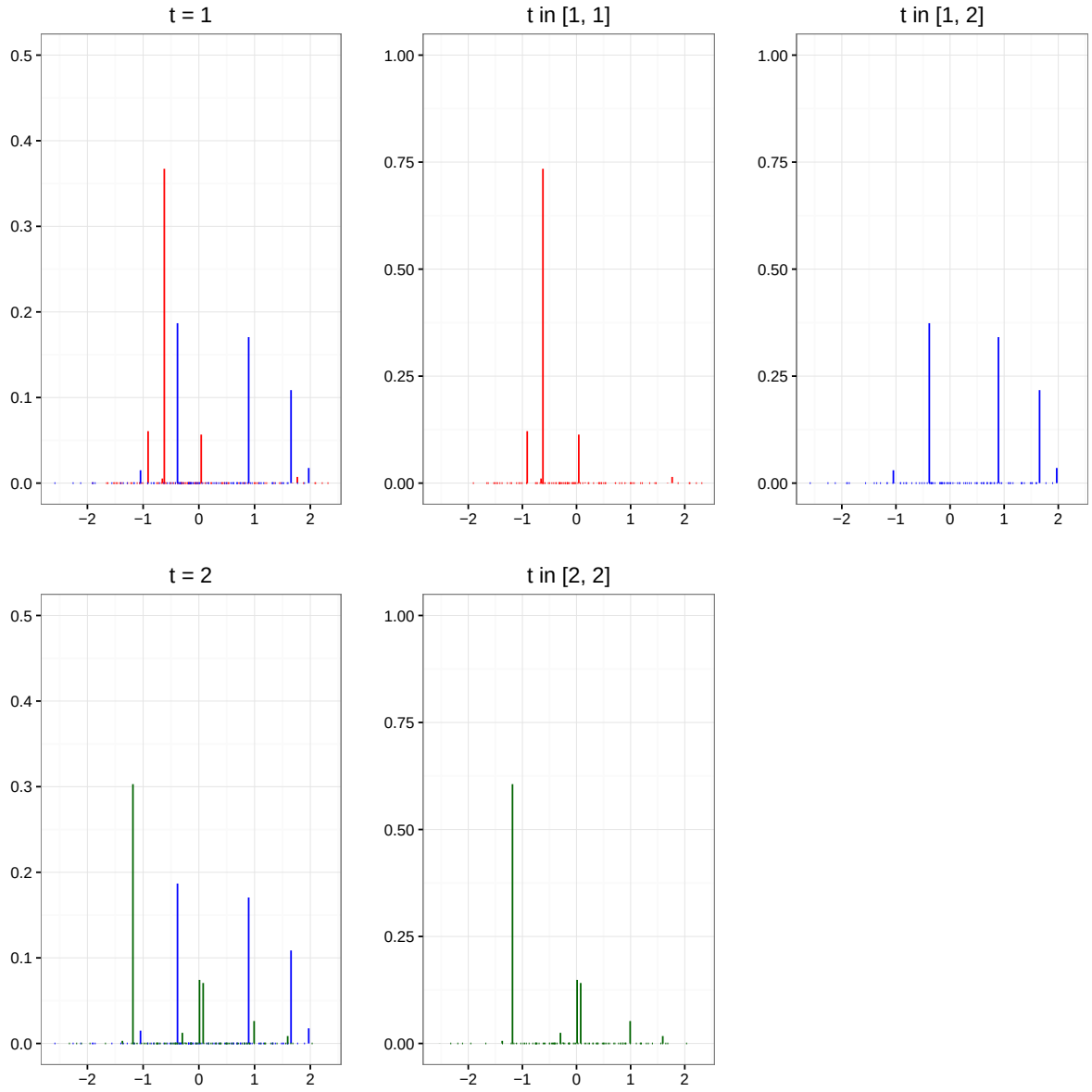

Figure 1: Illustration of mixing across contribution partitions in the SNΓP model. The leftmost column shows the Dirichlet processes obtained from mixing all appropriate contributions for each of two time points. The remaining columns show the constituent Gamma processes (here normalized) over the partitions R which contribute to the mixtures.

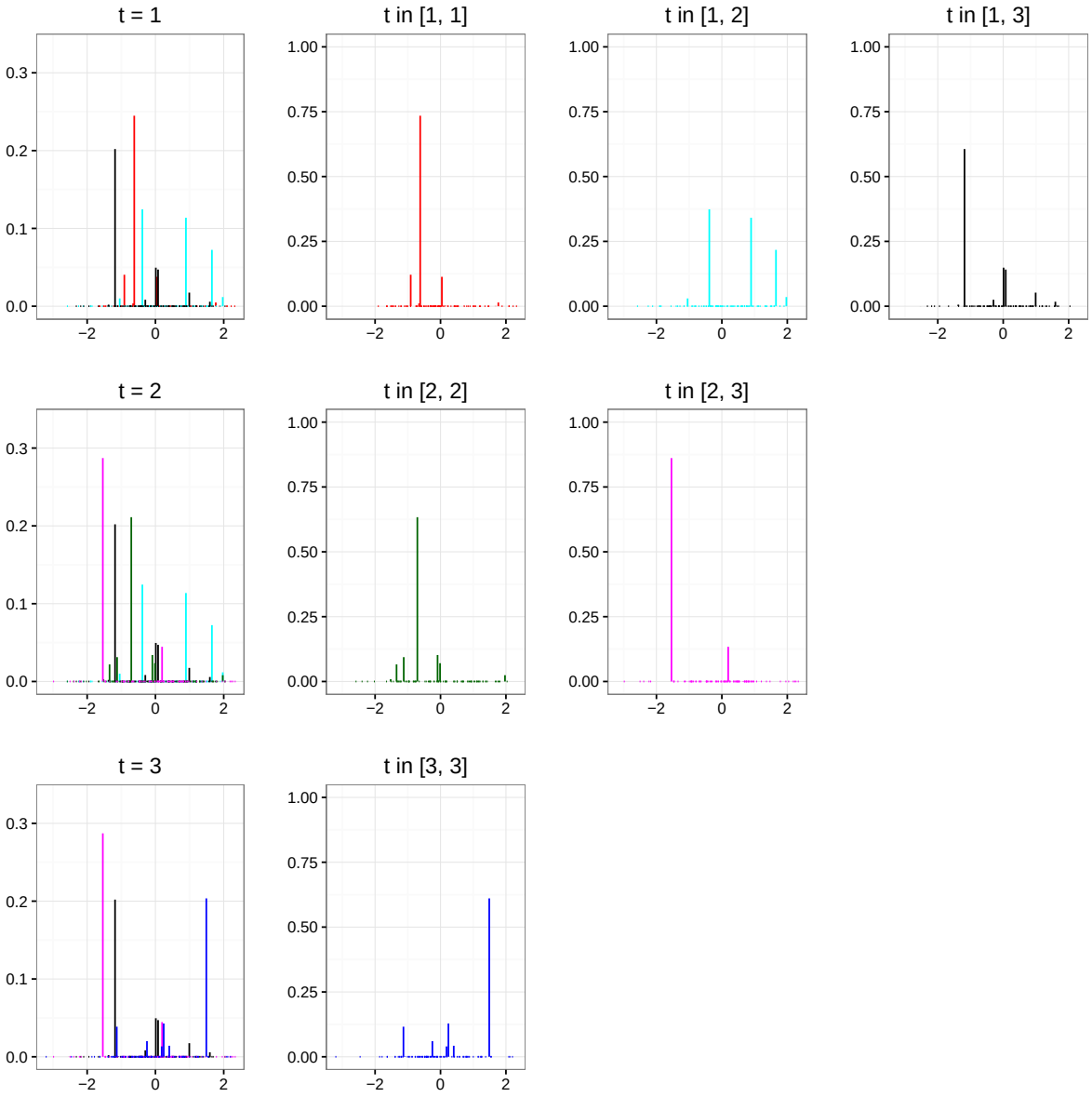

Figure 2: Illustration of SNΓP mixtures for three time points.

3.2 Dependence via stick-breaking

One possibility for introducing dependence into non-parametric models, is to do this via the stick-breaking construction. Recall that in general, the stick lengths (i.e. atom weights) π_k are obtained by $\pi_k = V_k \prod_{i<k}(1-V_i)$, where V_k are drawn i.i.d. from some distribution. To introduce dependence, we allow V_k depend on x, e.g. we may replace these with some stochastic process $V_k^{(x)}$. Next, we will explore several ways for doing this.

The *kernel stick-breaking process* [1] is based on the idea that if two observations are close in the covariate space $\mathcal{X}$, their stick-breaking ratios $V_k^{(x)}$ should be similar. Therefore, each atom θ_k is associated to a location in the covariate space μ_k, and $V_k^{(x)} := U_k K(x, \mu_k)$ where $K(\cdot, \cdot)$ is the kernel function and U_k are sampled i.i.d. from a Beta distribution. A common choice for K is the Gaussian kernel with a suitable bandwidth.

The *logistic stick-breaking process* [7] builds upon this idea, using a kernel logistic regression for each break of the stick. That is, $\log\left(\frac{V_k^{(x)}}{1-V_k^{(x)}}\right)$ is modeled by a weighted sum of kernels $K(x, \mu_i)$. In principle, the logistic link function is not the only choice and e.g. probit could be used.

Instead of modeling the $\text{logit}(V_k^{(x)})$ or $\text{probit}(V_k^{(x)})$ via kernels centered at some locations μ_k, one could introduce a latent underlying process to model this. The dependent *probit stick-breaking process* [8] introduces a latent Gaussian process for this purpose, i.e. it specifies $V_k^{(x)} := \Phi(\alpha_k^{(x)})$, where $\alpha_k^{(x)}$ is a Gaussian process for each k. Common choices for the covariance function $\gamma(x_i, x_j)$ include the exponential $\sigma^2 \exp(-||x_i - x_j||/\lambda)$ and Gaussian $\sigma^2 \exp(-||x_i - x_j||^2/\lambda)$ covariance functions. So, the strength of dependence between two nearby locations x_i and x_j is modeled via α_k and its behaviour is determined by our choice of covariance. Then, the Gaussian CDF Φ transforms this to interval $(0, 1)$, which results in dependent stick lengths.

The probit stick-breaking process based on a latent Gaussian process has been illustrated in Figure 3: realisations of the latent processes $\alpha_k^{(x)}$ are shown on the left, and the respective stick lengths $\pi_k^{(x)}$ in the middle. The random measures at three fixed time points are shown on the right. As seen from the figure, the dependency induced by this model differs from the one by SNTP. In Figure 3(b), the atom with the largest weight at time $t = 0$ decreases for time points $t \approx 0.5$, but for $t \approx 1$ it becomes dominating again. This differs from SNTP where the atoms are shared by nearby regions, and once the two time points become far apart, the atoms may no longer be shared.

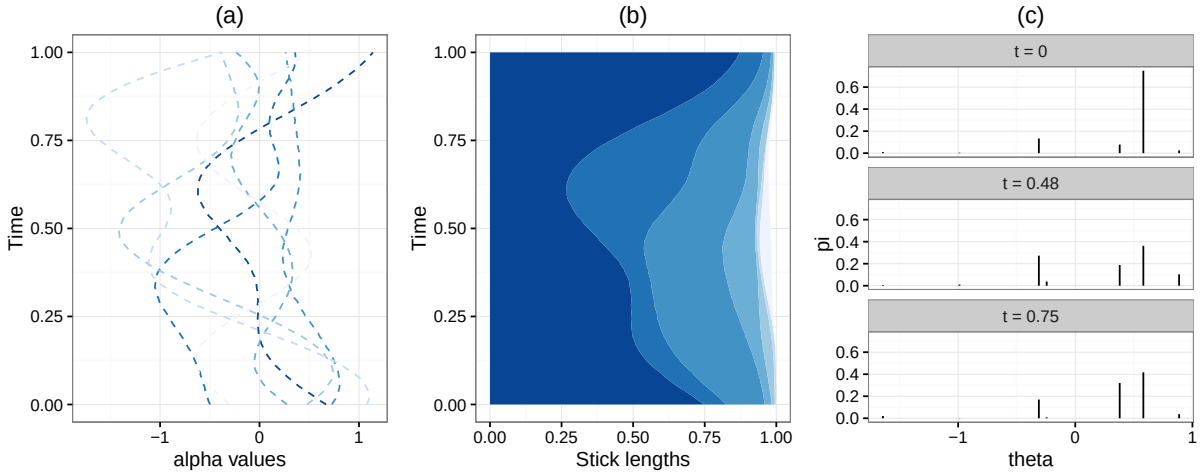

Figure 3: Example generated from the dependent probit stick-breaking process. The dependent covariate x here is time. (a) The latent processes α_k (we used mean 0 and covariance function $\gamma(t_i, t_j) = \exp\{-10(t_i - t_j)^2\}$ for this Gaussian Process). (b) The result of the probit stick-breaking process (stick lengths x-axis), based on the α for each time point (y-axis). (c) The obtained random measures at three time points ($t \in \{0, 0.48, 0.75\}$).

References

[1] David B Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

[2] Nicholas J Foti and Sinead A Williamson. A survey of non-exchangeable priors for Bayesian

nonparametric models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):359–371, 2015.

[3] J. F. C. Kingman. Uses of exchangeability. *Ann. Probab.*, 6(2):183–197, 04 1978.

[4] Steven N MacEachern. Dependent Dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 2000.

[5] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[6] Vinayak Rao and Yee W Teh. Spatial normalized gamma processes. In *Advances in neural information processing systems*, pages 1554–1562, 2009.

[7] Lu Ren, Lan Du, Lawrence Carin, and David Dunson. Logistic stick-breaking process. *The Journal of Machine Learning Research*, 12:203–239, 2011.

[8] Abel Rodriguez and David B Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian analysis (Online)*, 6(1), 2011.