Journal of Statistical Planning and Inference 146 (2014) 56–69

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A transformation approach to modelling multi-modal diffusions

Julie Lyng Forman [a,*], Michael Sørensen [b]

[a] Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, DK-1014 Copenhagen K, Denmark
[b] Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

ARTICLE INFO

Article history:
Received 3 April 2013
Received in revised form
6 September 2013
Accepted 26 September 2013
Available online 7 October 2013

Keywords:
Diffusion
Measurement error
Martingale estimating function
Multi-modality
Protein folding
Stochastic differential equation

ABSTRACT

This paper demonstrates that flexible and statistically tractable multi-modal diffusion models can be attained by transformation of simple well-known diffusion models such as the Ornstein–Uhlenbeck model, or more generally a Pearson diffusion. The transformed diffusion inherits many properties of the underlying simple diffusion including its mixing rates and distributions of first passage times. Likelihood inference and martingale estimating functions are considered in the case of a discretely observed bimodal diffusion. It is further demonstrated that model parameters can be identified and estimated when the diffusion is observed with additional measurement error. The new approach is applied to molecular dynamics data in the form of a reaction coordinate of the small Trp-zipper protein, from which the folding and unfolding rates of the protein are estimated. Because the diffusion coefficient is state-dependent, the new models provide a better fit to this type of protein folding data than the previous models with a constant diffusion coefficient, particularly when the effect of errors with a short time-scale is taken into account.

1. Introduction

In this paper we propose a new class of stationary stochastic differential equation models that have multi-modal invariant distributions. These models are useful for modelling dynamical systems that switch randomly between two or more regimes. As an example, we consider molecular dynamics data in the form of a protein reaction coordinate with two regimes corresponding to the folded and unfolded states of the protein, respectively. However, applications of bimodal diffusions are not limited to the study of molecular dynamics. Other applications of bimodal diffusion are as models of the global climate where the two regimes could be a cold and a hot climate as in Imkeller and Pavlyukevich (2002), and as financial models of e.g. interest rates subject to changes in the underlying financial and economic mechanisms as in Aït-Sahalia (1996).

Traditionally bimodal diffusion processes have been constructed by a stochastic differential equation with additive noise for a process moving in a double-well potential, i.e. a stochastic differential equation of the form

$$dY_t = -V'(Y_t)\,dt + \sigma^2\,dB_t, \tag{1}$$

where $\{B_t\}$ is a Wiener process and V is a potential with two valleys. Under the condition that $V(y)$ goes to infinity at the boundaries of the state space and that the function $h(y) = \exp\{-2V(y)/\sigma^2\}$ is integrable on the state space, $\{Y_t\}$ is ergodic

* Corresponding author. Tel.: +45 35327919.
E-mail address: jufo@sund.ku.dk (J.L. Forman).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.09.013

with invariant density being proportional to $h(y)$. An often studied example is given by the potential $V(y)=\theta y^2(y^2-2)$ with $\theta>0$, for which the drift is $-4\theta(y^3-y)=-4\theta y(y+1)(y-1)$. This simple potential has wells of the same depths at 1 and -1 and is symmetric around a separating potential barrier at 0. The related diffusion is ergodic with invariant density being proportional to $\exp\{-2\theta(y^4-2y^2)\}$. It is easy to generalize this model to models with wells at other points that need not be symmetric around the separating potential barrier. The double well potential models are the state of the art in the analysis of protein reaction coordinates such as the one considered in our case study. That is, constant diffusion is usually assumed and used in the estimation of the protein folding rates, see e.g. Socci et al. (1996). A more complex model of molecular dynamics was presented in Pokern et al. (2009) who used a partially observed hypoelliptical diffusion to model the dihedral angle in a butane molecule. Still this model assumes a constant diffusion coefficient which may conflict with that of the data. More recently evidence of non-constant diffusion in protein reaction coordinates has been reported in several articles. Best and Hummer (2010) discuss these findings and their implications for the assessment of protein folding rates.

Our new class of bimodal diffusions is obtained by applying particular transformations to simple well-known diffusions such as the Ornstein–Uhlenbeck process or a general Pearson diffusion; see Forman and Sørensen (2008). This leads to diffusion models with nonlinear drift and non-constant diffusion coefficients that are still highly tractable both from a statistical and a computational point of view. A major point of this paper is that many properties of diffusions are preserved by transformation. These include stationarity, mixing properties, and distributions of first passage times. Also the eigenvalues of the infinitesimal generator of the diffusion are preserved by the transformation, and eigenfunctions transform in a straightforward way. This facilitates efficient approximate likelihood inference by means of e.g. the explicit martingale estimating functions proposed by Kessler and Sørensen (1999). In the rare cases where the likelihood function of a diffusion is explicitly known, this is also the case for any of its transformations. Similarly to the double well potential models, our new diffusion models allow for great flexibility in the modelling of the invariant marginal distribution. Thus, the new bimodal diffusion models provide a useful extension of the class of bimodal diffusion models.

An alternative to the stochastic differential equation approach is to model each regime separately and to let the shifts between the models be determined by an underlying finite-state process such as a hidden Markov model or a Markov state model. These models are widely employed as models of protein folding, see Prinz et al. (2011) for a review, although it is recognized that the models are inadequate in describing the more gradual transition between states which is the de facto behaviour of many proteins. Latent Markov state models are also very popular in financial and econometric applications, see Lange and Rahbek (2009) for an overview. However, a latent state model is too complex and difficult to interpret if what is observed is not two different dynamical systems, but is really the same dynamical system that just has the property that it can be in two different regimes. A multi-modal diffusion has local attraction points corresponding to its regimes and moves between them in a continuous and random way. Apart from the conceptual advantage and the simpler model, other advantages of a multi-modal diffusion over two separate models are that the stationary marginal density in a succinct way contains important information about the regimes: the relative size of the modes reflect the time spent in each mode, and the peakedness and broadness of the modes reflect the volatility of the regimes. Moreover, explicit formulae for mean passage times allow for precise calculation of the time spent in each regime and the probability of switching from one regime to another.

The paper is organized as follows: Section 2 is devoted to the construction of our multi-modal diffusion models. We investigate the properties of these models and contrast them to other existing bimodal diffusions, the double-well potential models in particular. In Section 3 we discuss inference for the new class of bimodal diffusions emphasizing approximate likelihood inference based on martingale estimating functions. Inference is further discussed when the bimodal diffusion is observed with measurement error. In Section 4 we apply a bimodal diffusion model to molecular dynamics data in the form of a reaction coordinate of the small Trp-zipper protein. Upon adjusting for measurement error we obtain a good fit to data and estimated folding rates that are realistic for this kind of protein. Section 5 concludes.

2. Multi-modal diffusions by transformation

In order to model a bimodal (multi-modal) diffusion, we initially consider a stationary diffusion of general form,

$$dX_t=\mu(X_t)\,dt+\sigma(X_t)\,dB_t, \tag{2}$$

where $\{B_t\}$ is a Wiener process, and we assume that the coefficients are sufficiently regular to ensure that a unique weak solution exists for any given initial condition. In principle $\{X_t\}$ could be any diffusion, but we aim to construct bimodal diffusions for which statistical inference is relatively easy, so we are interested in cases where the diffusion $\{X_t\}$ is analytically tractable. This is for instance the case if $\{X_t\}$ is an ergodic Pearson diffusion as considered by Forman and Sørensen (2008), see in addition Kolmogorov (1931) and Wong (1964) for early accounts on these processes.

Recall that a stationary solution $\{X_t\}$ to the stochastic differential equation exists if

$$\int_{x^{hash;}}^{r} s(x)\,dx=\int_{\ell}^{x^{hash;}} s(x)\,dx=\infty \quad \text{and} \quad \int_{\ell}^{r}[s(x)\sigma^2(x)]^{-1}\,dx<\infty.$$