Probability Approximations with Applications in Computational Finance and Computational Biology

Yichao Wu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2006

Approved by:

Advisor: Chuanshu Ji
Advisor: Harry Hurd
Reader: Edward Carlstein
Reader: Ross Leadbetter
Reader: Yufeng Liu

© 2006
Yichao Wu
ALL RIGHTS RESERVED

ABSTRACT
YICHAO WU: Probability Approximations with Applications in Computational Finance and Computational Biology.
(Under the direction of Chuanshu Ji and Harry Hurd.)

In this work, certain probability approximation schemes are applied to two different contexts: one under stochastic volatility models in financial econometrics and the other about the hierarchical clustering of directional data on the unit (hyper)sphere. In both cases, approximations play an important role in improving the computational efficiency.

In the first part, we study stochastic volatility models. As an indispensable part of Bayesian inference using MCMC, we need to compute the option prices for each iteration at each time. To facilitate the computation, an approximation scheme is proposed for numerical computation of the option prices based on a central limit theorem, and some error bounds for the approximations are obtained.

The second part of the work originates from studying microarray data. After preprocessing the microarray data, each gene is represented by a unit vector. To study their patterns, we adopt hierarchical clustering and introduce the idea of linking by the size of a spherical cap. In this way, each cluster is represented by a spherical cap. By studying the distribution of direction data on the unit (hyper)sphere, we can assess the significance of observing a big cluster using Poisson approximations.

ACKNOWLEDGMENTS

I am deeply indebted to my advisors, Professors Chuanshu Ji and Harry Hurd, who guided this work and helped whenever I was in need. Without their constant support and encouragement, this work would not be possible.

I am also grateful to the faculty members in the Department of Statistics and Operations Research for their support; especially to my committee members Edward Carlstein, Ross Leadbetter, and Yufeng Liu.

Last but not the least, I would like to thank my family for their support. Without them, I would not have gone so far.

Contents

List of Figures vii

List of Tables viii

1 Introduction 1

2 Error Bounds for Gaussian Approximations of Option Prices in Stochastic Volatility Models 3

2.1 Introduction . . . 3

2.2 Basic framework and main results . . . 5

2.2.1 SV Models . . . 5

2.2.2 European call options pricing formulas and their discrete approximations . . . 7

2.2.3 Main result . . . 9

2.3 Methods . . . 12

2.4 Conclusion . . . 21

3 Clustering of Directional Data (with application to microarray data) 23

3.1 Microarrays and clustering . . . 23

3.2 Introduction to clustering . . . 24

3.2.1 Dissimilarity Measure . . . 25

3.2.2 Linkage . . . 26

3.2.3 Process of hierarchical clustering . . . 28

3.2.4 Dendrogram . . . 29

3.3 Clustering of directional data 31
3.3.1 Directional data from microarray data 31
3.3.2 Transformation from spherical polar coordinates to cartesian coordinates 33
3.3.3 Linking by size of spherical cap 34
3.3.4 Probability related to spherical cap assuming uniformity 35
3.4 Tests for uniformity 36
3.5 Detection of significant clusters in presence of an inhomogeneous Poisson background 39
3.6 Simulation 44
3.7 Summary 48

4 Summary of Dissertation **49**

A Proofs of Theorems and Lemmas **51**

Bibliography **70**

List of Figures

3.1 Assorted linking methods 27
3.2 Clustering 100 in unit square 30
3.3 One typical kernel smooth density of HG-U133plus gene expression. 32
3.4 Log-Log plot of node size vs linking distance. 38
3.5 Empirical CDF vs theoretical CDF of inter-event distance 39
3.6 True density of simulation 45
3.7 Significance test result for $\gamma = 0.5$ 47

List of Tables

3.1 WinBUGS estimator of mixture model 42

3.2 Distinguishing result of different cases 46

Chapter 1

Introduction

In this work, certain probability approximation schemes are applied to two different contexts: one under stochastic volatility models in financial econometrics and the other about the hierarchical clustering of directional data on the unit (hyper)sphere. In both cases, approximations play an important role in improving the computational efficiency.

More explicitly, in Chapter 2 we study stochastic volatility models. As an indispensable part of Bayesian inference using MCMC, we need to compute the option prices for each iteration at each time. To facilitate the computation, an approximation scheme is proposed for numerical computation of the option prices based on a central limit theorem, and some error bounds for the approximations are obtained.

The second part of this work originates from the study of microarray data. It is very common for biologists to ask the following questions: how does a gene react to different experimental conditions and what is the pattern? To answer these questions, in chapter 3 we study directional data on the unit (hyper)sphere, obtained from pre-processing the microarray data. We introduce the idea of hierarchical clustering based on the size of a spherical cap. Under this framework, each cluster is represented by a mean direction and a central angle that defines the boundary of a spherical cap, and those genes falling into the same cluster respond in a similar pattern to different experimental conditions. Hence, by studying the distribution of directional data, we can assess the significance of observing

a big cluster. Here in order to calculate the p-value, we adopt the Poisson approximation method.

Chapter 2

Error Bounds for Gaussian Approximations of Option Prices in Stochastic Volatility Models

2.1 Introduction

Stochastic volatility (SV) models become an important model family that extend the Black-Scholes (BS) theory in financial economics and adapt it to many applications in more realistic financial markets, such as option pricing and risk management. See the review paper by Ghysels, Harvey, and Renault (1996), and the references therein. The success in this area depends on our capability of calibrating SV models using financial data from two sources: underlying asset returns and option prices. Adding option data to this scenario creates a significant challenge for the following reason. In a SV model, a volatility time series consists of latent variables h_t, $t = 0, 1, \ldots, T$, where h_t is a function of the time-t volatility through a logarithmic or power transformation. Calibration of the series $\{h_t\}$ and the parameter θ (often multidimensional) involved in a SV model is an indirect inference problem, requiring some Monte Carlo based algorithms, such as efficient methods of moments (EMM) or Markov chain Monte Carlo (MCMC). There is

an extensive literature in calibration of SV models using both returns and option data. Here we only mention a few. See Chernov and Ghysels (2000), Pan (2002) for the EMM approach, and Eraker (2004), Ge (2000), Jones (2003), Cheng, Gallant, Ji, and Lee (2005) for the MCMC approach. Those algorithms yield updated values $\theta^{(m)}$ and $h_t^{(m)}$, $t = 0, 1, \ldots, T$ in the mth iteration, $m = 1, \ldots, M$. In particular, as an important part of the mth iteration in fitting a SV model, an option price $f_t^{(m)}$ is calculated by using a pricing formula based on the SV model with the current parameter value $\theta^{(m-1)}$ and volatility value $h_t^{(m-1)}$, and compared to the observed option price C_t. The comparison will lead to the adjustment of $h_t^{(m-1)}$ to $h_t^{(m)}$, and also the updating of $\theta^{(m-1)}$ to $\theta^{(m)}$ based on $h_t^{(m)}$, $t = 0, 1, \ldots, T$. Notice that a general method of computing $f_t^{(m)}$ is to perform high-dimensional numerical integration — treated as a conditional expectation over the space of sample paths of future volatility, and this has to be done for *every* t and *every* m in the algorithm. The required computational time quickly adds up with large values of T (several hundred days) and M (typically more than $100,000$ iterations). So far, there have been basically two ways to handle the computation in practice. A basic strategy is to use "brute force" simulation, i.e. to generate a large number of "additional" sample paths of future volatility based on which the Monte Carlo integration yields an approximation to $f_t^{(m)}$. An alternative approach relies on the availability of closed-form option pricing formulas which can avoid the aforementioned brute force numerical integration. However, Heston's model [see Heston (1993)] appears to be the only case beyond the original BS setting that enjoys a closed-form solution. The intractability of the brute force calculation and the limitation with closed-form pricing formulas present an urgent need for developing new efficient computational methodology.

In this work, we present an approximation scheme for numerical computation of the option price based on a central limit theorem (CLT), and provide some error bounds for the approximations. The proposed Gaussian approximations promote a significant dimension reduction for numerical integration, from the space of volatility sample paths (with

dimensionality of several hundreds) to the sample space of bivariate Gaussian vectors.

The basic framework and main result are presented in Section 2.2. Section 2.3 outlines the method for proving the main result along with all lemmas and propositions. Section 2.4 concludes. Appendix contains technical proofs.

2.2 Basic framework and main results

This section describes the SV models under objective and risk-neutral measures respectively, sets forth option pricing formulas, and states the result of Gaussian approximations for the option price with related error bounds.

2.2.1 SV Models

Let $S = \{S_t\}$ denote a continuous-time process that describes the evolution of an asset price. We presume that its logarithm follows the SV model:

$$
\begin{aligned}
y_t &= \log(S_t), && (2.1)\\
dy_t &= \mu\, dt + e^{h_t/2}\left(\sqrt{1-\rho^2}\, dW_{1t} + \rho\, dW_{2t}\right), && (2.2)\\
dh_t &= (\alpha + \beta h_t)\, dt + \sigma\, dW_{2t}, && (2.3)
\end{aligned}
$$

where μ is a deterministic drift and $W = \{W_t\} = \{(W_{1t}, W_{2t})\}$ is a standard 2D Wiener process defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_t : 0 \le t \le T\}$ denote the filtration generated by W. The parameter ρ is the correlation of the asset return and the volatility factor. This model is analogous to the discrete-time logarithmic first order autoregressive [AR(1)] SV model in Jacquier, Polson, and Rossi (2003).

Equivalent martingale measures

To price an option on S we follow a standard approach: We assume no arbitrage and specify two risk premia processes. In doing so, we follow established convention [Pan

(2002), Jones (2003), Polson and Stroud (2003), Eraker (2004), etc.] and specify the risk premia such that stochastic differential equations (SDEs) describing the returns process have the same functional form under the physical and risk neutral measures. These two risk premia processes are

$$\begin{aligned} \nu_t &= \nu_1 + \nu_2 h_t, && (2.4) \\ \lambda_t &= \frac{1}{\sqrt{1-\rho^2}} \left\{ e^{-h_t/2} \left[(\mu - r) + e^{h_t/2} \right] - \rho \nu_t \right\}, && (2.5) \end{aligned}$$

where r is the short rate, presumed constant. Assuming $E\left[\exp\left(\int_0^T (\lambda_u^2 + \nu_u^2)\, du\right)\right] < \infty$, the risk neutral valuation principle as shown in Harrison and Kreps (1979) and Harrison and Pliska (1981) implies that there exists an equivalent martingale measure Q such that the time-t price of a contingent claim $g(S_T)$ is expressed as $E^Q\left[e^{-r(T-t)} g(S_T) \mid \mathcal{F}_t\right]$ where $E^Q(\cdot)$ is the expectation operator under measure Q.

For the purpose of computing the above conditional expectation, the measure Q can be expressed either as the Radon-Nikodým derivative with respect to the physical measure P, which is

$$\xi_t = \left.\frac{dQ}{dP}\right|_{\mathcal{F}_t} = \exp\left(- \int_0^t \lambda_u \, dW_{1u} - \int_0^t \nu_u \, dW_{2u} - \frac{1}{2}\int_0^t \lambda_u^2 \, du - \frac{1}{2}\int_0^t \nu_u^2 \, du \right) \quad (2.6)$$

in this instance, or as SDEs representing the asset evolution under Q,

$$\begin{aligned} dS_t &= r S_t \, dt + e^{h_t/2} S_t \left(\sqrt{1-\rho^2} \, dW_{1t}^Q + \rho \, dW_{2t}^Q \right), && (2.7) \\ dh_t &= \left[\alpha - \nu_1 \sigma + (\beta - \nu_2 \sigma) h_t \right] \, dt + \sigma \, dW_{2t}^Q, && (2.8) \end{aligned}$$

where $W^Q = \{(W_{1t}^Q, W_{2t}^Q)\}$ is a standard 2D Wiener process under Q, given by

$$W_{1t}^Q = W_{1t} + \int_0^t \lambda_u du \quad (2.9)$$

$$W_{2t}^{Q} = W_{2t} + \int_0^t \nu_u du. \tag{2.10}$$

2.2.2 European call options pricing formulas and their discrete approximations

Let $\theta = \{\mu, \alpha, \beta, \sigma, \rho\}$ and $\nu = \{\nu_1, \nu_2\}$ denote the parameters of the SV model described in (2.1)–(2.3) and (2.7)–(2.8). Following Romano and Touzi (1997), for maturity (expiration) date T and strike price K, an European call option price can be expressed as

$$\begin{aligned}
\mathcal{C}(S_t, h_t, \theta, \nu) &= E^{Q}\left[C^{BS}\left(S_t\ e^{Z_{t,T}}, \overline{V}_{t,T}, r, K, T-t\right)|\mathcal{F}_t\right], &(2.11)\\
Z_{t,T} &= \rho \int_t^T e^{h_u/2}\ dW_{2u}^{Q} - \frac{1}{2}\ \rho^2\ \int_t^T e^{h_u}\ du, &(2.12)\\
\overline{V}_{t,T} &= \frac{1-\rho^2}{T-t} \int_t^T e^{h_u}\ du, &(2.13)
\end{aligned}$$

where the integration is taken over future volatility and $C^{BS}(\cdot)$ is the Black-Scholes call option pricing formula with the expression

$$\begin{aligned}
&C^{BS}(s, v, r, k, d) &(2.14)\\
&= s\Phi\left(\frac{\log(s/k) + (r + v^2/2)\, d}{v\sqrt{d}}\right) - ke^{-rd}\Phi\left(\frac{\log(s/k) + (r + v^2/2)\, d}{v\sqrt{d}} - v\sqrt{d}\right),
\end{aligned}$$

where Φ is the standard normal cumulative distribution function. Notice that with a constant volatility and $\rho = 0$, (2.11) returns to the original Black-Scholes call option price. A put option can be priced using the parity relation between a put (P_t) and a call (C_t), which is $P_t = C_t - S_t + K\ e^{-r(T-t)}$.

An obvious (and usual) approach to computing the integrals in (2.11)–(2.13) is to discretize (2.7)–(2.8), generate sample paths according to the recursions implied by the

discretization, evaluate (2.12)–(2.13) over each sample path, and average. Indeed, this is how we check the accuracy of our proposed Gaussian approximations to (2.11). For most purposes, this straightforward approach is too computationally intensive to be practicable. Therefore, a Gaussian approximation scheme was proposed in Cheng, Gallant, Ji, and Lee (2005) as a viable alternative approach.

Letting Δ be a small time increment, the discretizion of the volatility process under the risk-neutral measure Q, i.e. (2.8), by means of an Euler scheme [See Kloeden and Platen (1992), p341] yields the recursion

$$h_{t+\Delta} = a + b\ h_t + c\ \epsilon_{t+\Delta}, \tag{2.15}$$

where for each t, the innovation $\epsilon_{t+\Delta} \sim N(0,1)$ is independent of the spot volatility h_t (which summarizes the past history up to time t due to the Markovian property), and also independent of the three parameters: the intercept $a = (\alpha - \nu_1\sigma)\Delta$, the slope $b = 1 + (\beta - \nu_2\sigma)\Delta$, and the diffusion coefficient $c = \sqrt{\Delta}\sigma$. Notice that our notation departs from convention because we have incorporated Δ into the expressions for a, b, and c in order to simplify later formulas. Let $n = (T-t)\Delta^{-1}$. Our task is to derive the asymptotic joint distribution of the two sums

$$U_n = \sum_{j=0}^{n-1} e^{h_{t+j\Delta}}, \tag{2.16}$$

$$V_n = \sum_{j=0}^{n-1} e^{h_{t+j\Delta}/2}\ \epsilon_{t+(j+1)\Delta}, \tag{2.17}$$

considering the approximations of the two integrals, $U_n\ \Delta \approx \int_t^T e^{h_u}\ du$ and $V_n\ \sqrt{\Delta} \approx \int_t^T e^{h_u/2}\ dW_{2u}^Q$, that appear in (2.12)–(2.13).

Before stating our main result, here is an outline for how the proposed Gaussian approximation scheme works. The problem of interest is to compute the option price in (2.11)–(2.13). Our proposed method includes several steps and various approximations.

Step 1 Write

$$\begin{aligned} \mathcal{C}(S_t, h_t, \theta, \nu) &= E^Q \left[C^{BS} \left(S_t \, e^{Z_{t,T}}, \overline{V}_{t,T}, r, K, T-t \right) | \mathcal{F}_t \right] \\ &\approx E_t \, \widetilde{G} \left(U_n, V_n \right) \end{aligned} \tag{2.18}$$

for some function $\widetilde{G}$, where "$\approx$" reflects the approximation by Euler discretization of the risk-neutral dynamics of (2.7)–(2.8) using *small* Δ.

Step 2 Write

$$E_t \, \widetilde{G} \left(U_n, V_n \right) \approx E_t \, \widetilde{G} \left((Var_t U_n)^{1/2} U + E_t U_n, \; (Var_t V_n)^{1/2} V + E_t V_n \right), \tag{2.19}$$

where "$\approx$" denotes a practical use of the Gaussian approximations that should be justified by Theorem 1 and Theorem 2 (in section 2.2.3) with *fixed* Δ and *large* n. Here "practical use" and "justification" are two related but different aspects. In particular, the function G in Theorem 2 is bounded and its expression does not depend on n, while the function $\widetilde{G}$ need not be bounded and its expression does involve n.

Step 3 Compute $E_t \, \widetilde{G} \left((Var_t U_n)^{1/2} U + E_t U_n, \; (Var_t V_n)^{1/2} V + E_t V_n \right)$ by simulating the 2D Gaussian vector (U, V) or using some numerical integration methods (not Monte Carlo) such as Gaussian quadratures.

2.2.3 Main result

For completeness we quote Theorem 1 [from Cheng, Gallant, Ji, and Lee (2005)] which concerns a CLT for (U_n, V_n) and gives explicit expressions for the asymptotic conditional means, variances and covariance. Theorem 2, as a refinement of Theorem 1, provides an upper bound for the errors incurred when applying Theorem 1 to the calculation of the

European call option price $\mathcal{C}(S_t, h_t, \theta, \nu)$.

For fixed t, let $E_t(\cdot)$, $Var_t(\cdot)$ and $Cov_t(\cdot)$ denote the conditional expectation, variance and covariance operators respectively, given h_t and under Q.

Theorem 1 *Assume $|b| < 1$. Fix t and an arbitrary initial state h_t. As $n \to \infty$, the limiting distribution of $n^{-1/2}(U_n - E_t U_n,\ V_n - E_t V_n)$ conditioning on h_t is a bivariate normal distribution with mean $(0,0)$ and covariance matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ provided $a_{11}a_{22} > a_{12}^2$, with*

$$\begin{aligned} a_{11} &= \lim_{n\to\infty} n^{-1} Var_t(U_n), \\ a_{12} &= a_{21} = \lim_{n\to\infty} n^{-1} Cov_t(U_n, V_n), \\ a_{22} &= \lim_{n\to\infty} n^{-1} Var_t(V_n); \end{aligned}$$

where

$$E_t U_n = \sum_{i=0}^{n-1} \exp\left[\frac{a(1-b^i)}{1-b} + b^i h_t + \frac{c^2(1-b^{2i})}{2(1-b^2)}\right]; \tag{2.20}$$

$$E_t V_n = 0; \tag{2.21}$$

$$Var_t(U_n) = \sum_{i=0}^{n-1} Var_t(e^{h_{t+i\Delta}}) + 2\sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1} Cov_t(e^{h_{t+i\Delta}}, e^{h_{t+j\Delta}}); \tag{2.22}$$

$$\begin{aligned} Cov_t(U_n, V_n) = \sum_{i=1}^{n-1}\sum_{j=0}^{i-1} c\, b^{i-j-1} \exp\Bigg[\left(b^i + b^j/2\right) h_t + \frac{a\left(3/2 - b^i - b^j/2\right)}{1-b} \\ + \frac{c^2\left(5/4 - b^{2i} - b^{2j}/4 + b^{i-j} - b^{i+j}\right)}{2(1-b^2)}\Bigg]; \end{aligned} \tag{2.23}$$

$$Var_t(V_n) = E_t U_n; \tag{2.24}$$

with

$$
\begin{aligned}
& Var_t(e^{h_{t+i\Delta}}) \\
= \; & \exp\left(2b^i h_t + \frac{2a(1-b^i)}{1-b} + \frac{c^2(1-b^{2i})}{1-b^2}\right)\left[\exp\left(\frac{c^2(1-b^{2i})}{1-b^2}\right) - 1\right],
\end{aligned}
$$

$$
\begin{aligned}
& Cov_t(e^{h_{t+i\Delta}}, e^{h_{t+j\Delta}}) \\
= \; & \exp\left((b^i + b^j)h_t + \frac{a(2-b^i-b^j)}{1-b} + \frac{c^2(2-b^{2i}-b^{2j})}{2(1-b^2)}\right) \\
& \cdot \left[\exp\left(\frac{c^2(b^{j-i} - b^{j+i})}{1-b^2}\right) - 1\right].
\end{aligned}
$$

Note: Without loss of generality, we set $t = 0$ and define the normalized sums

$$
\begin{aligned}
U_n^{(0)} &= (Var_0 U_n)^{-1/2}(U_n - E_0 U_n) \\
V_n^{(0)} &= (Var_0 V_n)^{-1/2}(V_n - E_0 V_n).
\end{aligned}
$$

It is equivalent to Theorem 1 that $\left(U_n^{(0)},\ V_n^{(0)}\right)$ conditioning on h_0 converges to (U, V) in distribution as $n \to \infty$, where (U, V) follows a bivariate normal distribution with mean $(0, 0)$ and covariance matrix $\begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}$. Note that $\rho_0 \in [-1, 1]$ does not depend on the initial state h_0 since $\{h_t\}$ is an ergodic Markov chain. But h_0 will affect the convergence rate and error bounds given in the following theorem.

Theorem 2 *Fix h_0 and Δ, and assume $|b| < 1$. For any bounded C^2 function $G(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$, there exist constants $B > 0$, $\delta \in (0, 1)$ such that*

$$
\left|E_0 G\left(U_n^{(0)},\ V_n^{(0)}\right) \ - \ E_0 G(U, V)\right| \ \leq \ B\ n^{-\delta} \tag{2.25}
$$

for all sufficiently large n.

Theorem 1 follows from a CLT for strong mixing sequences. The calculation needed for the conditional means, variances and covariance, although tedious, is mostly straightforward. However, the argument required for Theorem 2 is much more challenging. In the end, the upper bounds (represented by the coefficient B and exponent δ) we derived are not desirably sharp. The numerical range of δ is determined by a number of constraints in Lemma 3. Putting those constraints in a simple numerical optimization algorithm yields $\delta \approx 0.07$ — an unsatisfactory small value. The limitation seems partly due to the method for proving adopted in this work. See Section 2.3 for the discussion on related method for proving. Nevertheless, the numerical studies conducted in Cheng, Gallant, Ji, and Lee (2005) have demonstrated great promise with the proposed Gaussian approximation scheme, i.e. the actual errors incurred by using the approximations are quite small, and in the meantime, the computational intensity is significantly reduced.

2.3 Methods

To derive upper bounds for $\left|E_0 G\left(U_n^{(0)},\ V_n^{(0)}\right)\ -\ E_0 G(U,V)\right|$ is a classical problem in probability. The required method is dictated by the structure of (U_n, V_n). In our case, (U_n, V_n) is represented as an additive functional of the AR(1) process h. Since h is an ergodic Markov chain, a preferred method is to apply the spectral theory of Markov transition operators (or infinitesimal operators), i.e. the spectral gap usually provides sharp upper bounds for the convergence rates. For arbitrary $u, v \in I\!R$, write $uU_n + vV_n = \sum_{j=0}^{n-1} \xi_j$ where $\xi_j = ue^{h_{j\Delta}} + ve^{h_{j\Delta}/2}\epsilon_{(j+1)\Delta}$. The difficulty is that each additive term ξ_j involves not just a single component $h_{j\Delta}$, but a pair $h_j^{(2)} = (h_{j\Delta}, h_{(j+1)\Delta})$ which also forms a Markov chain. However, the transition probability operator $L_{h^{(2)}}$ for $h^{(2)} = \{h_j^{(2)},\ j = 0, 1, 2, ...\}$ is not a (strict) contraction. More specifically, let L_h be the transition probability operator associated with h and π the stationary distribution of h [we also

denote the density for π simply by $\pi(x)$]. Consider

$$\lambda = \sup \frac{\|L_h\ g\|}{\|g\|}, \tag{2.26}$$

where the sup is over all complex-valued functions g on $\mathbb{R}$ with $E^\pi g = 0$, and $\|\ \|$ is the $L^2(\mathbb{R}, \pi)$ norm:

$$\|g\|^2 = \int_R |g(x)|^2 \pi(x) dx.$$

Under the assumption $|b| < 1$ in (2.15), the operator L_h is a contraction because $\lambda < 1$ in (2.26). Note that the Markov chain $h^{(2)}$ has the following properties:

(i) The state space is $\mathbb{R}^2$ but the transition density $f_{h^{(2)}}(\cdot|(x,y))$ has a 1D support, i.e. given $(x,y) \in \mathbb{R}^2$,

$$f_{h^{(2)}}((x',y')|(x,y)) = \begin{cases} f_h(y'|y) & \text{if } x' = y, \\ 0 & \text{if } x' \neq y, \end{cases}$$

where $f_h(\cdot|x)$ is the transition density for h.

(ii) $h^{(2)}$ is ergodic and has the unique stationary distribution $\pi^{(2)}$ with density $\pi^{(2)}(x,y) = \pi(x) f_h(y|x)$.

(iii) The transition probability operator $L_{h^{(2)}}$ is not a contraction. To verify this, let g be a nonconstant function such that $E^\pi g = 0$. Define $g^{(2)}(x,y) = g(x)$ then $g^{(2)}$ is a nonconstant function on $\mathbb{R}^2$ and

$$E^{\pi^{(2)}} g^{(2)} = E^\pi g = 0;$$

and

$$\begin{aligned} L_{h^{(2)}}\ g^{(2)}(x,y) &= \int_R g^{(2)}(y,z)\ f_{h^{(2)}}((y,z)|(x,y))\ dz \\ &= \int_R g(y)\ f_h(z|y)\ dz \\ &= g(y). \end{aligned}$$

Hence

$$\begin{aligned} \|g^{(2)}\|^2 &= \int_{R^2} |g^{(2)}(x,y)|^2 \pi(x) f_h(y|x)\ dxdy \\ &= \int_R |g(x)|^2 \pi(x)\ dx \\ &= \int_R |g(y)|^2 \pi(y)\ dy \end{aligned}$$

$$\begin{aligned} &= \int_{R^2} |g(y)|^2 \pi(x) f_h(y|x)\ dxdy \\ &= \int_{R^2} |L_{h^{(2)}}\ g^{(2)}(x,y)|^2 \pi(x) f_h(y|x)\ dxdy \\ &= \|L_{h^{(2)}}\ g^{(2)}\|^2. \end{aligned}$$

Thus the function $g^{(2)}$ is orthogonal to all constant functions, and not strictly contracted by the operator $L_{h^{(2)}}$.

An alternative approach, also the one adopted in this work, is based on the strong mixing property of AR(1) processes. There is an extensive literature in CLT and related convergence rates for stationary processes with mixing properties. Since the option price is represented as a conditional expectation given the present asset price and relevant information, the underlying AR(1) process is a non-stationary process starting from a fixed initial value. We first consider the stationary case. Using the classical splitting technique [often referred to as the "big-block-small-block" (BBSB) method], we derive

error bounds for characteristic functions in Gaussian approximations and convert them to error bounds for approximating option prices as bounded functionals. This part is a modification of the results in Bhattacharya and Rao (1976), and in Reznik (1968). To obtain error bounds for the ultimate non-stationary case, we apply the coupling method to link the AR(1) process starting from a fixed state (the non-stationary case) to the AR(1) process starting from its stationary distribution π (the stationary case). The exponential mixing rate for AR(1) processes implies an exponential tail probability for the coupling time. The coupling time is defined rigorously for Ornstein-Uhlenbeck (O-U) processes [see Øksendal (1995) for general definition] — a continuous-time counterpart of AR(1) processes. The arguments in this part move back and forth between AR(1) processes and O-U processes at our convenience. This approach often reduces the sharpness of error bounds due to the limitation that the sizes of big/small blocks in the splitting technique increase with n, which prevents us from utilizing detailed information within each block in the asymptotics.

Here is a summary of lemmas and propositions that lead to Theorem 2.

- Lemma 1: an exponential mixing rate for AR(1) process
- Lemma 2: asymptotic first and second moments of $(U_n,\ V_n)$ (the stationary case)
- Lemma 3: an error bound for approximating the characteristic function of the standardized sums $(U_n^*,\ V_n^*)$ (the stationary case)
- Proposition 1: an error bound for approximating the call option price expressed as an expectation of bounded function of $(U_n^*,\ V_n^*)$ (the stationary case)
- Lemma 4: an exponential tail probability of the crossing time (or coupling time) for two AR(1) processes
- Theorem 2: an error bound for approximating the call option price expressed as an expectation of bounded function of $(U_n^{(0)},\ V_n^{(0)})$ (the non-stationary case)

Notation: Discretizing h with time increment Δ and observing h only at $t = k\Delta$, $k = 0, 1, ..., n$, we simply write h_k for $h_{k\Delta}$, hence the AR(1) process (2.15) is expressed as

$$h_{k+1} = a + b\ h_k + c\ \epsilon_{k+1}, \quad k = 0, 1, ..., n. \tag{2.27}$$

Let $\sigma(X)$ denote the σ-field generated by random variable (or random vector) X; $f_X(\cdot)$ the density of X; $f_{Y|X}(\cdot|x)$ the conditional density of Y given $X = x$; and $f(x; \mu, \sigma)$ the normal density for $N(\mu, \sigma)$.

Recall the definition of strong mixing given in Bradley (1986). Let $\{X_i,\ i = 0, 1, 2, \ldots\}$ be a (discrete time) stochastic process, and $\mathcal{F}_m^n$ the σ-algebra generated by $X_i,\ m \le i \le n$. $\{X_i\}$ is said to be strong mixing with rate α_n if

$$\alpha_n = \sup_{t \ge 0} \sup_{A \in \mathcal{F}_0^t,\ B \in \mathcal{F}_{t+n}^{\infty}} |P(A \cap B) - P(A)P(B)| \longrightarrow 0$$

as $n \to \infty$.

Lemma 1 *Assume $|b| < 1$. Then the AR(1) process h in (2.27) is strong mixing with an exponential mixing rate,*

$$\alpha(n) < c_1\ e^{-b_1 n} \tag{2.28}$$

for some constant $c_1 > 0$, where $b_1 = -\log|b| > 0$.

The following lemma is the stationary counterpart of Theorem 1 in which the initial state h_0 follows the stationary distribution π for h.

Lemma 2 *Assume $h_0 \sim N\left(\frac{a}{1-b}, \frac{c^2}{1-b^2}\right)$ and $|b| < 1$. Let*

$$\begin{aligned} U_n &= \sum_{j=0}^{n-1} e^{h_j}, \\ V_n &= \sum_{j=0}^{n-1} e^{h_j/2} \epsilon_{j+1}. \end{aligned}$$

Then we have

$$
\begin{aligned}
EU_n &= n \ \exp\left[\frac{a}{1-b}+\frac{c^2}{2(1-b^2)}\right]; && (2.29)\\
EV_n &= 0; && (2.30)\\
Var(U_n) &= \exp\left[\frac{2a}{1-b}+\frac{c^2}{1-b^2}\right] \\
&\quad \left[n\left(\exp(\frac{c^2}{1-b^2})-1\right)+2\ \sum_{j=1}^{n-1}(n-j)\left(\exp(\frac{c^2b^j}{1-b^2})-1\right)\right]; && (2.31)\\
Var(V_n) &= EU_n; && (2.32)\\
Cov(U_n,V_n) &= \exp\left[\frac{3a}{2(1-b)}+\frac{5c^2}{8(1-b^2)}\right] \\
&\quad \sum_{j=1}^{n-1}(n-j)\ c\ b^{j-1}\ \exp\left[\frac{c^2b^j}{2(1-b^2)}\right]. && (2.33)
\end{aligned}
$$

Hence

$$
\begin{aligned}
\sigma^2_{U_n} &= Var(U_n) = n\ \sigma_U^2(1+o(1)) = O(n), && (2.34)\\
\sigma^2_{V_n} &= Var(V_n) = n\ \sigma_V^2(1+o(1)) = O(n), && (2.35)\\
&\quad \textit{for some positive constants } \sigma_U \textit{ and } \sigma_V \textit{ as } n\to\infty, \textit{ and} \\
\lim_{n\to\infty} corr\ (U_n,\ V_n) &= \rho_0. && (2.36)
\end{aligned}
$$

Define the normalized sums

$$
\begin{aligned}
U_n^* &= \sigma_{U_n}^{-1}\sum_{j=0}^{n-1} x'_j, && (2.37)\\
V_n^* &= \sigma_{V_n}^{-1}\sum_{j=0}^{n-1} x''_j, && (2.38)
\end{aligned}
$$

where $x'_j = e^{h_j} - \exp(\frac{a}{1-b}+\frac{c^2}{2(1-b^2)})$ and $x''_j = e^{h_j/2}\epsilon_{j+1}$. The next lemma (Lemma 3) concerns an error bound of the Gaussian approximation for the characteristic function of $(U_n^*,\ V_n^*)$ (with the Cramér-Wold device), which sets the stage for developing other error

bounds that will eventually lead to the result of Theorem 2.

Lemma 3 *Assume* $h_0 \sim N\left(\frac{a}{1-b}, \frac{c^2}{1-b^2}\right)$, $|b| < 1$ *in (2.27). For* $\sqrt{u^2+v^2} \leq n^{\varrho}$ *with some* $\varrho \in (0,1)$ *and* $i = \sqrt{-1}$,

$$\begin{aligned} & \left| E \exp[i\ (uU_n^* + vV_n^*)] - \exp\left[\frac{-1}{2}(u^2 + 2\rho_0 uv + v^2)\right] \right| \\ = \ & O(k\alpha(q)) + \exp\left[\frac{-1}{2}(u^2 + 2\rho_0 uv + v^2)\right] \\ & \cdot \ \left[(u^2+v^2)\ O(n^{-\epsilon}) + (|u|+|v|)^3\ O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon})\right] \\ + \ & \sqrt{O(k\alpha(q)) + O((u+v)^2 n^{-2\epsilon}) + O((u+v)^3 n^{\delta_3 - 2\delta_3\epsilon - 2\epsilon - 1/2})}, \end{aligned} \tag{2.39}$$

where $\epsilon \in (0, 1/2)$, $\delta_2 \in (0,1)$, $\delta_3 \in (0,1)$ *and* $\varrho \in (0,1)$ *satisfy*

$$\begin{aligned} 2\varrho + 1/2 + \epsilon - 1 &< 0 \\ 3\varrho + (1+2\delta_2)(1/2+\epsilon) - 3/2 &< 0 \\ 2\varrho - 1/2 - \epsilon &< 0 \\ 3\varrho + (1+2\delta_3)(1/2-\epsilon) - 3/2 &< 0. \end{aligned}$$

To convert error bounds for characteristic functions to error bounds for bounded continuous functionals of cdf's, we extend the arguments given in Bhattacharya and Rao (1976) for the iid case to the case of strong mixing sequences.

First, we state a basic result in Fourier analysis. For any $f \in L^1(\mathbb{R}^2)$, denote its Fourier transform by

$$\hat{f}(u,v) = \int_{R^2} e^{i(ux_1+vx_2)} f(x_1,x_2)\ dx_1 dx_2, \quad (u,v) \in \mathbb{R}^2.$$

Let $K(\cdot)$ be a probability measure on $\mathbb{R}^2$ whose density $k(\cdot)$ is given by

$$k(x) = g_{a',4}(x_1)\ g_{a',4}(x_2), \tag{2.40}$$

where $x = (x_1, x_2)$, $a' = 2\pi^{-1/3}2^{5/6}$ and $g_{a',4}(y) = \frac{3a'}{2\pi}\ (\frac{\sin a'y}{a'y})^4$ with $y \in I\!R$. Note that

$$\begin{aligned}\hat{K}(u,v) &= \hat{g}_{a',4}(u)\ \hat{g}_{a',4}(v) = 0 \\ &\quad \text{if} \quad (u,v) \notin [-4a', 4a']^2 = \{(u,v) : |u| \le 4a', |v| \le 4a'\},\end{aligned}$$

where $(u,v) \in I\!R^2$. And

$$K(\{x :\ |x| \ge 1\}) \le \frac{6a'}{\pi} \int_{\sqrt{2}/2}^{\infty} \left(\frac{\sin a'y}{a'y}\right)^4 dy \le 1/8.$$

Hence

$$K(\{x :\ |x| < 1\}) \ge 7/8. \tag{2.41}$$

For function g and $\varepsilon > 0$, define

$$\begin{aligned}\omega_g(A) &= \sup\{|g(x) - g(y)| :\ x, y \in A\} \quad \text{where} \quad A \subset I\!R^2, \\ \bar{\omega}_g(\varepsilon : \mu) &= \int \omega_g(B(x,\varepsilon))\mu(dx) \quad \text{where} \quad B(x,\varepsilon) = \{y :\ |y - x| \le \varepsilon\}, \\ \omega_g^*(\varepsilon : \mu) &= \sup\{\bar{\omega}_{g_y}(\varepsilon : \mu) : y \in I\!R^2\}\end{aligned}$$

where μ is a finite measure and $g_y(x) = g(y + x)$.

Proposition 1 *Assume* $h_0 \sim N\left(\frac{a}{1-b}, \frac{c^2}{1-b^2}\right)$, $|b| < 1$ *in (2.27), and* $g(\cdot,\cdot) :\ I\!R^2 \to I\!R$ *is a bounded* C^2 *function with bounded first-order derivatives. Let* Q_n *denote the cdf for* (U_n^*, V_n^*) *and* $\Phi^{(2)}$ *the (limiting) cdf for* (U, V)*. Then we have*

$$\begin{aligned}& |Eg(U_n^*, V_n^*) - Eg(U,V)| \\ = &\ \left|\int_{R^2} g\ d(Q_n - \Phi^{(2)})\right| \\ \le &\ O(n^{4\varrho-\epsilon}) + O(n^{2\varrho+(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon}) \\ + &\ O(n^{7\varrho/2+(\delta_3-2\delta_3\epsilon-2\epsilon-1/2)/2}) + O(n^{-\gamma_2}) + O(n^{-\varrho}) + \omega_g^*(8\sqrt{2}a'n^{-\varrho} : \Phi^{(2)})\end{aligned} \tag{2.42}$$

where $a' = 2\pi^{-1/3}2^{5/6}$, $\epsilon \in (0, 1/2)$, $\delta_2 \in (0,1)$, $\delta_3 \in (0,1)$, $\varrho \in (0,1)$ *and* $\gamma_2 \in (0,1)$ *[taken from Lemma 1 in Reznik (1968)] satisfy*

$$\begin{aligned} 2\varrho + 1/2 + \epsilon - 1 &< 0 \\ 3\varrho + (1+2\delta_2)(1/2+\epsilon) - 3/2 &< 0 \\ 2\varrho - 1/2 - \epsilon &< 0 \\ 3\varrho + (1+2\delta_3)(1/2-\epsilon) - 3/2 &< 0. \end{aligned}$$

Having obtained the error bounds for the stationary case, we will consider the non-stationary case in which the initial state h_0 in h is arbitrarily given. The method of coupling will allow us to make connections between the two cases. For convenience, we will move back and forth between AR(1) time series and their continuous-time counterpart — O-U processes. Note that the solution of SDE (2.8) is an O-U process. We assume $\beta - \nu_2\sigma < 0$ which is consistent with the assumption $|b| < 1$ in the discrete-time AR(1) model (2.15) or (2.27).

Lemma 4 *Suppose both* $h^{(1)}$ *and* $h^{(\pi)}$ *are embedded in the same (enlarged) probability space. They follow the same dynamics (2.15), driven by two independent innovation processes [both consisting of iid* $N(0,1)$ *components]. Let* $h^{(1)}$ *start from an arbitrary initial value* $h_0^{(1)} = z_1$, *while* $h_0^{(\pi)} \sim \pi$, *where* π *is the stationary distribution* $N\left(\frac{a}{1-b}, \frac{c^2}{1-b^2}\right)$. *Define the crossing time of these two chains by*

$$\tau(z_1) = \inf\{k \geq 1 : sign(h_k^{(\pi)} - h_k^{(1)}) \neq sign(h_0^{(\pi)} - z_1)\}. \tag{2.43}$$

Then for sufficiently large n, *we have*

$$P(\tau(z_1) > n) \leq c_1 \exp(-c_2 n^{1/2-\delta_1}) \tag{2.44}$$

where $c_2 = -\log(\min(|b|, \Phi(1)))$, Φ *is the cdf of* $N(0,1)$, c_1, δ_1 *are some positive constants, and* δ_1 *can be arbitrarily small.*

Using the same notation with subscript $t \geq 0$, let $\{h_t^{(1)}\}$ and $\{h_t^{(\pi)}\}$ be two independent (continuous-time) O-U processes that satisfy the same SDE (2.8) but have different initial states: $h_0^{(1)} = z_1$, and $h_0^{(\pi)} \sim N\left(\frac{a}{1-b}, \frac{c^2}{1-b^2}\right)$.

Definition 1 *Define the coupling time*

$$T_c(h^{(1)}, h^{(\pi)}) = \inf\{t \geq 0 : \ h_t^{(1)} = h_t^{(\pi)}\}; \tag{2.45}$$

and the coupled process $h_t^{(c)}$ *(the superscript "c" stands for coupling)*

$$h_t^{(c)} = \begin{cases} h_t^{(1)} & t < T_c(h^{(1)}, h^{(\pi)}) \\ h_t^{(\pi)} & t \geq T_c(h^{(1)}, h^{(\pi)}). \end{cases}$$

The following connection between the coupling time $T_c(h^{(1)}, h^{(\pi)})$ and the crossing time $\tau(h_0^{(1)})$ is obvious:

$$T_c(h^{(1)}, h^{(\pi)}) \leq \tau(h_0^{(1)})\ \Delta. \tag{2.46}$$

Note that $\{h_t^{(c)}\}$ and $\{h_t^{(1)}\}$ have the same distribution, so do their discretized versions. In fact, $\{h_t^{(c)}\}$ and $\{h_t^{(1)}\}$ can be considered versions of the same process in the enlarged probability space. See appendix for proofs.

2.4 Conclusion

In this chapter, we have obtained an upper bound for polynomial convergence rates of the proposed Gaussian approximation scheme. The argument we adopted limits the sharpness of the derived bound. More specifically, the discrete sums used to approximate the option price integrals involve a two-step Markov chain whose spectral gap is zero. Hence the

preferred method of spectral analysis for the Markov transition probability operator does not apply. Instead, we resort to the "big-block-small-block" splitting technique to deal with functionals of the underlying strong mixing processes.

Chapter 3

Clustering of Directional Data (with application to microarray data)

3.1 Microarrays and clustering

Humans have 20,000 to 25,000 genes (see Stein (2004)), each of which consists of a sequence of bases. These bases are the building blocks of DNA, which is often referred to as the molecule of heredity as it is responsible for the genetic propagation of most inherited traits. There are many techniques to study DNA. Among these, one of the most advanced techniques is the gene microarray, which measures the degree to which various genes are "expressed" in a sample. Recently, we have seen explosive use of DNA microarray in biological and medical research. Its application includes studying complex disease at the molecular level, detecting genes responsible for some clinical outcomes and many others.

By analyzing the changes in gene expression, scientists can study how cells respond to a disease or some external environmental challenge. But it is well known that microarrays produce overwhelming multivariate data. For example, the Affymetrix HGU133plus array contains 54,675 probe sets, representing 24,192 genes. Due to the size of the data, its analysis can be very complicated and thereby statistics plays an important role. The exact method that the analyst adopts depends much on the design of the experiment

being analyzed. Fold change (studying the ratio of expression of treatment condition over control) and relatively straightforward statistical analysis are appropriate in the case that there are only two samples (control and experiment) being compared. The analysis can get much more complex when there are more than two conditions. In this case, the same approach can be adopted pairwise, but more advanced multivariate analysis, clustering, can give more information.

For the huge datasets produced by DNA microarray, it is desirable to perform some sort of exploratory data analysis to study the responde pattern of gens with respect to experimental conditions. Clustering, one of the most important unsupervised learning methods, is very appropriate for this purpose. Its ultimate goal is to find a structure (not known a priori) in a set of unlabeled objects (sometimes called individuals, cases, data rows or observations) by forming clusters. These clusters are formed in such a way that the objects within a cluster are more similar to each other than objects assigned to other clusters. Hence clustering microarray data can help us to identify groups of genes that respond in a similar pattern to the assorted experimental conditions, see Eisen, Spellman, Brown, and Botstein (1998).

3.2 Introduction to clustering

Typically, in the clustering setting, we have n objects $\mathbf{x}_i$, $i = 1, 2, \cdots, n$ and each object $\mathbf{x}_i$ can be represented by a row vector in $\mathbf{R}^p$, i.e. $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$. Putting all together, we have a data matrix $\mathbf{X}$ of size $n \times p$, each row corresponding to an object. As an unsupervised learning, clustering aims at partitioning the objects into homogenous groups. In fact, there are many clustering methods, and they can be classified as either partitioning methods or hierarchical methods.

In partitioning methods, we assume that there are a fixed number of clusters and want to assign each object into a cluster according to some criteria (say to minimize the

total within-cluster dissimilarity). They include K-means MacQueen (1967), K-medoids Kaufman and Rousseeuw (1990), and Self-organizing maps Kohonen (1990). On the other hand, hierarchical clustering does not assign all the objects into clusters together in a single step. Instead, it proceeds stagewise producing a nested sequence of partitions, each of which corresponds to a different number of clusters. Technically, hierarchical clustering can be further divided into "agglomerative" (also known as bottom-up method), meaning that clusters are merged, and "divisive" (top-down method), in which each stage involves splitting one or more groups. Here we are going to focus on more popular agglomerative hierarchical clustering.

Agglomerative hierarchical clustering produces a nested sequence of partitions of the objects: P_n, P_{n-1}, $\cdots$, P_1. In particular, the first partition P_n, at the beginning, consists of n single-object clusters, and the last one P_1, in the end, is exactly a single cluster containing all the n objects. At each stage, the method merges together the two clusters which are closest together or most similar to each other. It is the so called linkage method.

Next, to measure the closeness between object(s) and cluster(s), we introduce dissimilarity distance and linkage. They are two of the most fundamental concepts in hierarchical clustering.

3.2.1 Dissimilarity Measure

There are many measures of dissimilarity appropriate for clustering. Here we give a brief list of the most commonly used ones. (Remark that any (semi-)metric $d(\mathbf{x}, \mathbf{y})$ can serve as a dissimilarity measure.)

- **Euclidean distance**

$$d(\mathbf{x}_i, \mathbf{x}_j) = \| \mathbf{x}_i - \mathbf{x}_j \|,$$

where $\| \mathbf{x} \| = \sqrt{\sum_{k=1}^{p} x_k^2}$ denotes the usual norm of the vector $\mathbf{x} = (x_1, x_2, \cdots, x_p)$.

- **Chebychev distance**

$$d(\mathbf{x}_i, \mathbf{x}_j) = \max\{|x_{ik} - x_{jk}|, k = 1, 2, \cdots, p\}.$$

- **City block or Manhattan distance**

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|.$$

- **Cosine distance**

$$d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\| \mathbf{x}_i \| \| \mathbf{x}_j \|},$$

where $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{k=1}^{p} x_{ik} x_{jk}$ denotes the inner product of two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\| \mathbf{x}_i \| = \sqrt{\langle \mathbf{x}_i, \mathbf{x}_i \rangle}$. Here it is worthwhile to point out that the cosine distance is a scale invariant distance measure, i.e., $d(a\mathbf{x}_i, b\mathbf{x}_j) = d(\mathbf{x}_i, \mathbf{x}_j)$, for any $a, b > 0$. More on this later.

- **Correlation distance**

$$d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\langle \mathbf{x}_i - \bar{\mathbf{x}}_i, \mathbf{x}_j - \bar{\mathbf{x}}_j \rangle}{\| \mathbf{x}_i - \bar{\mathbf{x}}_i \| \| \mathbf{x}_j - \bar{\mathbf{x}}_j \|},$$

where $\bar{\mathbf{x}}_i = (\frac{1}{p} \sum_{k=1}^{p} x_{ik}, \frac{1}{p} \sum_{k=1}^{p} x_{ik}, \cdots, \frac{1}{p} \sum_{k=1}^{p} x_{ik})$ is referred as the mean vector of $\mathbf{x}_i$. We can easily see that the correlation distance is also scale invariant.

3.2.2 Linkage

Linkage is referred to as the method used to define the dissimilarity distance of two clusters and it includes:

- **Single linkage or nearest-neighbor method**

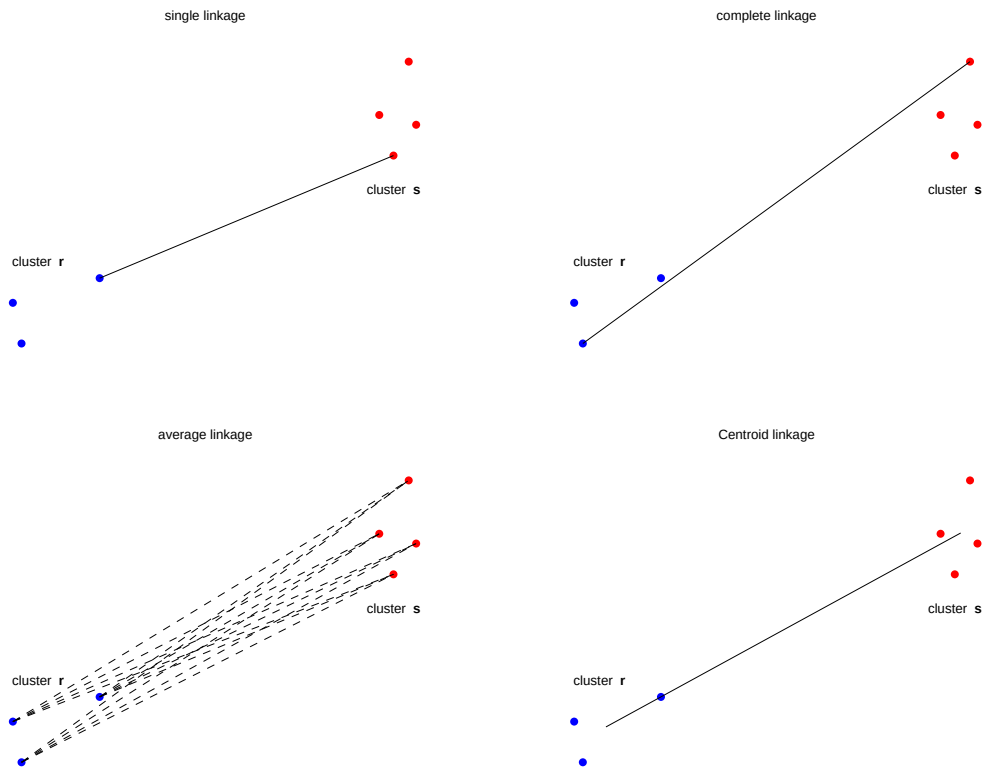

Figure 3.1: Assorted linking methods

Single linkage is one of the simplest methods to define the linking distance of two clusters. It defines the dissimilarity distance between two clusters **r** and **s** as follows,

$$D(\mathbf{r}, \mathbf{s}) = \min\{d(\mathbf{x}_i, \mathbf{x}_j) : \text{ object } i \in \mathbf{r} \text{ and object } j \in \mathbf{s}\}, \tag{3.1}$$

which is exactly the distance between the closest pair of objects, where only pairs consisting of one object from each cluster are considered.

- **Complete linkage or farthest-neighbor method** Opposite of the single linkage, the complete linkage defines inter-cluster distance as the distance between the farthest pair of objects (one from each cluster). Mathematically, it means as follows:

(and see the top-right panel of Figure 3.1 for illustration.)

$$D(\mathbf{r}, \mathbf{s}) = \max\{d(\mathbf{x}_i, \mathbf{x}_j) : \text{ object } i \in \mathbf{r} \text{ and object } j \in \mathbf{s}\}. \tag{3.2}$$

- **Average linkage** In average linkage clustering, the dissimilarity distance between clusters is defined to be the average distance between each pair of objects (one from each cluster). See left-bottom panel of Figure 3.1 for explanation.

$$D(\mathbf{r}, \mathbf{s}) = \text{Average}\{d(\mathbf{x}_i, \mathbf{x}_j) : \text{ object } i \in \mathbf{r} \text{ and object } j \in \mathbf{s}\}. \tag{3.3}$$

- **Centroid linkage** Centroid linkage defines the inter-cluster distance to be the distance between the centroids of the two clusters.

$$D(\mathbf{r}, \mathbf{s}) = d(\bar{\mathbf{x}}_\mathbf{r}, \bar{\mathbf{x}}_\mathbf{s}), \tag{3.4}$$

  where $n_\mathbf{r}$ denotes the number of objects in the cluster $\mathbf{r} = \{r_1, r_2, \cdots, r_{n_\mathbf{r}}\}$ and $\bar{\mathbf{x}}_\mathbf{r} = \frac{1}{n_\mathbf{r}} \sum_{k=1}^{n_\mathbf{r}} \mathbf{x}_{r_k}$ denotes the centroid of cluster $\mathbf{r}$. It is illustrated in the right-bottom panel of Figure 3.1.

In the sequel, hierarchical clustering in this manuscript would by default mean agglomerative hierarchical clustering.

3.2.3 Process of hierarchical clustering

With the preceding terminology for dissimilarity distance and linkage method, we are ready to introduce the detailed technique of hierarchical clustering. We begin with data matrix $\mathbf{X}$ of size $n \times p$ and there are n individual simple clusters, here by simple we mean that each cluster contains one and exactly one object. The dissimilarity distance measure and linking method are now fixed. At each stage of hierarchical clustering, we compare

the pairwise inter-cluster dissimilarity distance $D(\cdot, \cdot)$ among all current clusters, and find the minimizer pair, say the pair of clusters **r** and **s** minimizes $D(\cdot, \cdot)$. Then we merge the clusters **r** and **s** into one new cluster, hence the number of current clusters decreases by 1. This process is repeated until there is only one cluster left which includes all the n objects. This is summarized in the following algorithm. See Johnson (1967).

Algorithm A: Algorithm of general hierarchical clustering

1. Begin with all the single clusters, and set the sequence number $m = 0$.

2. Find the least dissimilar pair of clusters among all the current clusters, say a pair of clusters **r** and **s** which minimize $D(\cdot, \cdot)$ over all pairs of current clusters.

3. Increase the sequence number by 1: $m = m + 1$. Merge clusters **r** and **s** to form a new cluster. Set the level of this clustering as: $L(m) = D(\mathbf{r}, \mathbf{s})$.

4. Update the dissimilarity matrix M_D, by deleting the rows and columns corresponding to cluster **r** and **s**, and adding a row and a column corresponding to the newly formed cluster. The dissimilarity distance between the new cluster and any other existing cluster is computed using the adopted linking method and hence the dissimilarity matrix M_D is updated correspondingly.

5. The algorithm terminates if all objects are in one cluster. Otherwise, go to step 2.

3.2.4 Dendrogram

When there are a moderate number of objects, it is desirable to visualize the implication of the clusters. A graphical tree diagram, called dendrogram, displays the results of clustering (or of the linking). It shows the objects, the sequence of the clusters, and the linking distance between the clusters. In a dendrogram, the horizontal axis displays the indices of the objects, while the vertical axis shows the linking distance between the clusters. Each leaf in the dendrogram tree represents one object, and each branch

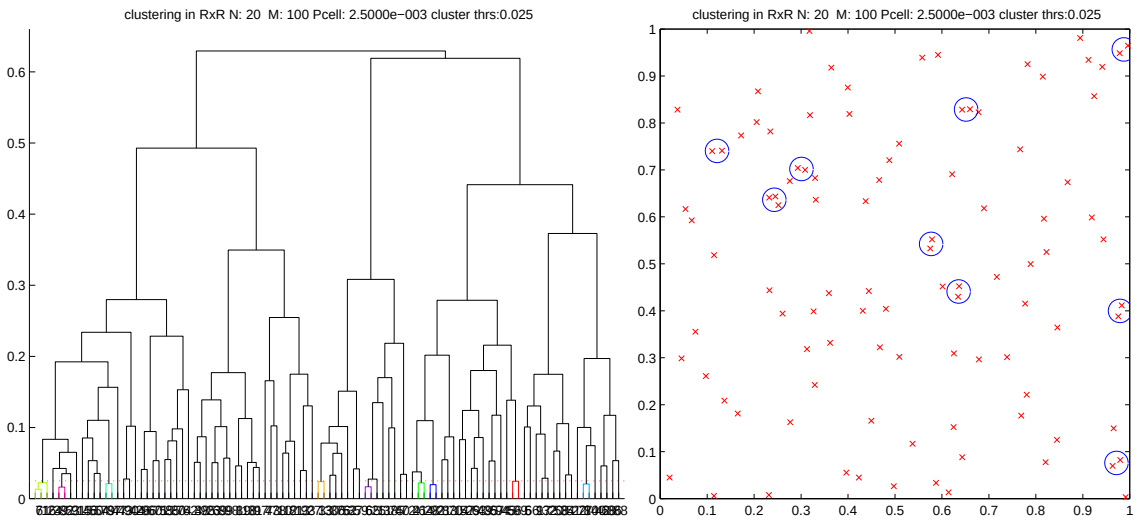

Figure 3.2: 100 points are uniformly distributed on unit square $[0,1] \times [0,1]$. Hierarchical cluster with Euclidean distance and average linkage is performed. The left panel is the resulting dendrogram with a cutting horizontal line at $d = 0.025$, and the right panel plots the 100 points and formed clusters in blue circles at this linking level.

denotes a cluster. If two clusters are connected, the height at which they merge is their linking dissimilarity distance. The lower (higher) two clusters are connected, the more (less) similar they are. Because of this special structure of dendrogram, we can study the behavior of the objects locally by cutting the dendrogram at a desirable level. After the cutting, we see many branches below the horizontal line corresponding to the cutting level, each representing a cluster. Some of these branches may degenerate to leaves, one-object clusters. All these branches together form a partition of all the objects.

Example 1: To give a better illustration, we simulate 100 data points uniformly distributed in the unit square $[0,1] \times [0,1]$. On this data set, we apply the hierarchical clustering using Euclidean distance and average linkage. The resultant dendrogram is plotted in the left panel of Figure 3.2. Further, we cut the dendrogram at the level of linking distance $d = 0.025$; the colorful branches denote the non-simple clusters (those with at least 2 points). The right panel of Figure 3.2 plots the 100 data points (each represented by a red cross sign) and the blue circles represents the non-simple clusters obtained by cutting the dendrogram at $d = 0.025$.

3.3 Clustering of directional data

3.3.1 Directional data from microarray data

In microarray experiments, we observe the expression value of each gene under each condition. In this case, each gene corresponds to an object and its corresponding vector is composed of the expression values of this gene under the assorted experimental conditions. Sometimes there may be more than one replicate for some condition(s). In this case, the corresponding entry in the data matrix is the mean expression value among all the replicates. One remarkable observation of the microarray data is that genes have a wide range of expression values. Figure 3.3 plots the kernel smooth density of the binary logarithms of the HG-U133plus gene expression for one experiment. We can see that it ranges from almost zero to as large as 16 after taking logarithm with base 2. Although the magnitude of expression differs a lot for different genes, it does not matter much for biological study because biologists are more interested in studying the relative changes from one condition to another. More precisely, they want to study the pattern of gene expressions across all the conditions, and thus identify groups of genes that respond in a similar way to these experimental conditions. Say, there are two genes: i and j; their expressions are proportional to each other, i.e. $\mathbf{x}_i = c\mathbf{x}_j$, for some $c > 0$. It is not hard for us to understand that they are similar in some way. For this consideration, it is unreasonable to use the Euclidean distance since they may be very far away from each other if $c > 0$ is large.

For the above reason and to study the pattern of gene expression, it is very natural to choose the correlation distance as follows,

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\langle \mathbf{x}_i - \bar{\mathbf{x}}_i, \mathbf{x}_j - \bar{\mathbf{x}}_j \rangle}{\| \mathbf{x}_i - \bar{\mathbf{x}}_i \| \| \mathbf{x}_j - \bar{\mathbf{x}}_j \|} = 1 - \left\langle \frac{\mathbf{x}_i - \bar{\mathbf{x}}_i}{\| \mathbf{x}_i - \bar{\mathbf{x}}_i \|}, \frac{\mathbf{x}_j - \bar{\mathbf{x}}_j}{\| \mathbf{x}_j - \bar{\mathbf{x}}_j \|} \right\rangle \tag{3.5}$$

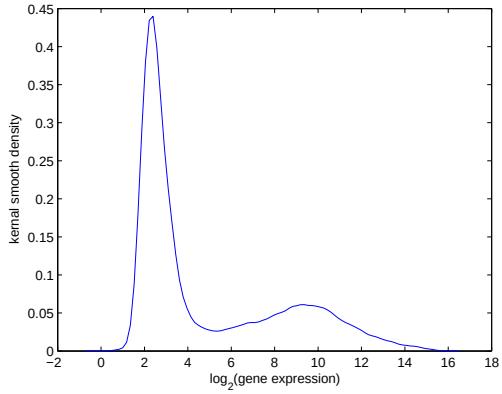

Figure 3.3: One typical kernel smooth density of HG-U133plus gene expression.

where $\bar{\mathbf{x}}_i = (\frac{1}{p}\sum_{k=1}^{p} x_{ik}, \frac{1}{p}\sum_{k=1}^{p} x_{ik}, \cdots, \frac{1}{p}\sum_{k=1}^{p} x_{ik})$ is the mean direction of $\mathbf{x}_i$.

Motivated by (3.5), we define *standardizing* as subtracting the mean vector and then normalizing to make the Euclidean length of the vector equal to one. That is, the standardized vector $\tilde{\mathbf{x}}_i^*$ of $\mathbf{x}_i$ is

$$\tilde{\mathbf{x}}_i^* = \frac{\mathbf{x}_i - \bar{\mathbf{x}}_i}{\| \mathbf{x}_i - \bar{\mathbf{x}}_i \|}. \tag{3.6}$$

Then with this notation, $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ can be expressed as $1 - \langle \tilde{\mathbf{x}}_i^*, \tilde{\mathbf{x}}_j^* \rangle$. We can easily see that each $\tilde{\mathbf{x}}_i^*$ is nothing but a unit vector on the unit p-sphere surface embedded in p-dimensional space, and hence are called *directional data*. In addition, because we subtract the mean direction $\bar{\mathbf{x}}_i$ here, $\tilde{\mathbf{x}}_i^*$ is orthogonal to direction $(1, 1, \cdots, 1)$ and thus in a manifold of dimension $p-1$. That is, choose a orthogonal transformation matrix M, whose last column corresponds to $\frac{1}{\sqrt{p}}(1, 1, \cdots, 1)^T$, then the last entry of $\tilde{\mathbf{x}}_i^* M$ is always zero. And denote the vector of the first $p-1$ elements of $\tilde{\mathbf{x}}_i^* M$ by $\tilde{\mathbf{x}}_i$. Hence $\tilde{\mathbf{x}}_i$ is on the unit $(p-1)$-sphere. Using this new notation, we can see that the correlation distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is exactly one minus the inner product of the corresponding vectors $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$, and thus exactly the cosine distance between these two vectors.

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = 1 - \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \rangle = d_{cos}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j).$$

So from now on, we are going to focus on directional data $\tilde{\mathbf{X}}$ by projecting each row $\mathbf{X}$ to the hyperplane orthogonal to direction $(1, 1, \cdots, 1)$, and then followed by standardizing. It deserves to point out that $\tilde{\mathbf{X}}$ is of size $n \times (p-1)$ if $\mathbf{X}$ is $n \times p$.

3.3.2 Transformation from spherical polar coordinates to cartesian coordinates

Each directional data on the unit p-sphere can be expressed using the spherical polar coordinates: $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_{p-1})$, $0 \leq \theta_i \leq \pi$, $i = 1, 2, \cdots, p-2$, $0 \leq \theta_{p-1} < 2\pi$, (here we do not have radius because each vector on the unit $p-1$-sphere has radius one) via the following transformation

$$\mathbf{x} = \mathbf{u}(\boldsymbol{\theta}),$$

where $\mathbf{u}_i(\boldsymbol{\theta}) = \cos\theta_i \prod_{j=0}^{i-1} \sin\theta_j, i = 1, 2, \cdots, p, \sin\theta_0 = \cos\theta_p = 1$.

If we denote the infinitesimal probability of directional data on the unit p-sphere S_p by dS_p, then we have

$$dS_p = a_p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where

$$a_p(\boldsymbol{\theta}) = \prod_{j=2}^{p-1} \sin^{p-j}\theta_{j-1}, \qquad \text{for } p > 2, \quad \text{and } a_2(\boldsymbol{\theta}) = 1.$$

To facilitate later simulation, here we introduce some methods to generate uniform directional data on the unit p-sphere. So far we have two methods to generate uniform distribution on the unit p-sphere.

1. Generate p-vector from a rotationally invariant distribution, such as $N(0, I_p)$, and normalize it to have unit length, where I_p is the $p \times p$ identity matrix.

2. Independently generate θ_i, $i = 1, 2, \cdots, p-2$, according to the following densities,

$$f(\theta_i) \propto \sin^{p-i}\theta_{i-1}, \quad 0 \leq \theta_i \leq \pi,$$

and θ_{p-1} from Uniform $[0, 2\pi)$. Then using above transformation $\mathbf{x} = \mathbf{u}(\boldsymbol{\theta})$, we get directional data $\mathbf{x}$ uniformly distributed on the unit p-sphere.

3.3.3 Linking by size of spherical cap

Usually linking is done according to a dissimilarity distance between clusters $\mathbf{r}$ and $\mathbf{s}$. Here we propose to link based on the size of the union of $\mathbf{r}$ and $\mathbf{s}$, where the size is defined as follows,

$$D_s(\mathbf{r}, \mathbf{s}) = \min_{\mathbf{x}_0 \in S_p} \max_{i \in \mathbf{r} \cup \mathbf{s}} \gamma(\mathbf{x}_i, \mathbf{x}_0), \tag{3.7}$$

where

$$S_p = \{\mathbf{z} \in \mathbf{R}^p : \| \mathbf{z} \| = 1\}$$

denotes the unit p-sphere in p-dimensional space. In the meanwhile, for $\mathbf{x}_0 \in S_p$, we also define a spherical cap centered at $\mathbf{x}_0$ with central angle θ as follows

$$S_{p,\theta}(\mathbf{x}_0) = \left\{ \mathbf{z} \in S_p : \cos\theta \leq \frac{\langle \mathbf{z}, \mathbf{x}_0 \rangle}{\| \mathbf{z} \| \| \mathbf{x}_0 \|} \leq 1 \right\} \tag{3.8}$$

Since the cosine distance is related to the central angle of spherical cap, the linking size can be defined in terms of the central angle of the smallest spherical cap to cover all the objects in the two clusters $\mathbf{r}$ and $\mathbf{s}$, that is,

$$\Theta(\mathbf{r}, \mathbf{s}) = \min_{\mathbf{x}_0 \in S_p} \left\{ \min_{\theta} \{\theta : \mathbf{x}_i \in S_{p,\theta}(\mathbf{x}_0), \forall i \in \mathbf{r} \cup \mathbf{s}\} \right\}. \tag{3.9}$$

One benefit of linking by size is that the linking value can tell us something about the probability assigned to a cap of this size.

With this special linkage (linking by size), we can apply **Algorithm A** to do hierarchical clustering on directional data. In this case, if we cut the dendrogram at certain dissimilarity distance level, each of the obtained clusters is represented by a spherical

cap. (But things can go wrong if you cut the dendrogram at a very high level because the obtained spherical caps may overlap.)

3.3.4 Probability related to spherical cap assuming uniformity

Suppose $\mathbf{x}_0$ is some arbitrary fixed vector on the unit p-sphere and $\mathbf{x}$ is random and *uniformly distributed* on S_p. Then the probability that $\mathbf{x}$ will be in the spherical cap $S_{p,\theta}(\mathbf{x}_0)$ is just the area of the cap relative to that of the whole unit sphere. That is,

$$\begin{aligned} \Pr[\mathbf{x} \in S_{p,\theta}(\mathbf{x}_0)] = \Pr[\langle \mathbf{x}, \mathbf{x}_0 \rangle > \cos\theta] &= \frac{\text{Area}\{S_{p,\theta}(\mathbf{x}_0)\}}{\text{Area}\{S_p\}} && (3.10) \\ &\triangleq \pi(\theta), && (3.11) \end{aligned}$$

where the last line just explicitly states that the probability is invariant with respect to $\mathbf{x}_0$ because $\mathbf{x}$ is uniformly distributed. By taking this one step further, assuming that $\mathbf{x}_0$ is random too, uniformly distributed on the unit p-sphere but independent of $\mathbf{x}$, we have

Proposition 2 *If the random vectors* $\mathbf{x}$, $\mathbf{y}$ *in* S_p *are independent and both uniformly distributed on* S_p, *then*

$$Pr[\langle \mathbf{x}, \mathbf{y} \rangle > \cos\theta] = \pi(\theta). \tag{3.12}$$

Proof: Let μ be the uniform measure on the Borel sets of S_p.

$$\begin{aligned} \Pr[\langle \mathbf{x}, \mathbf{y} \rangle > \cos\theta] &= \int \Pr[\langle \mathbf{x}, \mathbf{y} \rangle > \cos\theta | \mathbf{y}] \mu(d\mathbf{y}) \\ &= \int \pi(\theta) \mu(d\mathbf{y}) \\ &= \pi(\theta). \end{aligned}$$

3.4 Tests for uniformity

Here we seek to determine whether the observed clustering result is significant in some sense. A simple and natural one is to test whether the point pattern follows a homogeneous Poisson point process on the unit sphere, which is the same as to test whether the objects are uniformly distributed when the total number of points on the unit sphere is fixed.

$$\begin{aligned} H_0 &: \mathbf{x}_i\text{'s are uniformly distributed} \\ H_a &: \mathbf{x}_i\text{'s are not uniformly distributed} \end{aligned}$$

As pointed out earlier, cutting the dendrogram at a certain level generates a partition of the objects, which are a set of clusters each represented by a spherical cap. We can carry out this test by studying the resulting clusters. Say the largest number of points in these clusters is n_0, that is, there is at least one cluster containing n_0 points. Using the clumping heuristics in Aldous (1989), we can approximate the probability of observing such a large (or larger) cluster at this linking level under the null hypothesis. More precisely, we determine the probability that there exists some spherical cap containing at least n_0 points if there are n points uniformly distributed on the unit sphere S_p randomly.

Under the null hypothesis H_0, data points $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ on the unit p-sphere S_p follow a homogeneous Poisson process with estimated intensity $\lambda = \frac{n}{\text{Area}(S_p)}$. Let θ_K denote the central angle of the smallest spherical cap on S_p that contains K points. Then $\theta_K \leq \theta$ if and only if there exists a cluster of K or more points with cut $1 - \cos\theta$ in the dendrogram representation. Hence $P(\theta_K \leq \theta)$ can be considered as p-values for testing the null hypothesis H_0 of homogeneous Poisson. Aldous (1989) provides a guideline for approximating $P(\theta_K \leq \theta)$ under the assumption of small θ. Aldous (1989) only considers discs in $\mathbf{R}^2$, here we are going to generalize Aldous's argument to spherical caps in S_p to derive an approximation formula for $P(\theta_K \leq \theta)$. Fix a point $\mathbf{y}_0 \in S_p$ (say the "north pole"

$\mathbf{x}_{np} = (1, 0, \cdots, 0)$). Suppose there exist K points $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_K$ in the cap centered at $\mathbf{y}_0$ with central angle θ. Let

$$\zeta_{\theta,K} = \cap_{k=1}^{K} S_{p,\theta}(\mathbf{y}_k). \tag{3.13}$$

Let $a(p, \theta)$ denote the surface area of a spherical cap in S_p with central angle θ. We have

$$\Pr(\theta_K > \theta) \approx \exp\left\{-E\left[\frac{1}{\text{Area}(\zeta_{\theta,K})}\right] e^{-[\lambda a(p,\theta)]} \frac{[\lambda a(p,\theta)]^K}{K!}\right\}, \tag{3.14}$$

where $E\left[\frac{1}{\text{Area}(\zeta_{\theta,K})}\right]$ can be computed using Monte Carlo simulation.

To simulate $E\left[\frac{1}{\text{Area}(\zeta_{\theta,K})}\right]$, first we generate a large number of points (denoted by $\mathbf{Z}$) uniformly in $S_{p,2\theta}(\mathbf{y}_0)$, i.e. inside the spherical cap with central angle 2θ centered at $\mathbf{y}_0$. Then each time we generate K points $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_K$ uniformly within $S_{p,\theta}(\mathbf{y}_0)$, and calculate the fraction of the points $\mathbf{Z}$ falling inside the intersection of $S_{p,\theta}(\mathbf{y}_i)$'s, $i = 1, 2, \cdots, K$. Note that this fraction is approximately equal to $\text{Area}(\zeta_{\theta,K})/a(p, 2\theta)$. Hence we can estimate $E\left[\frac{1}{\text{Area}(\zeta_{\theta,K})}\right]$ by the empirical average of $\frac{1}{\text{Area}(\zeta_{\theta,K})}$ over many repetitions.

Example 2: Before introducing another test, we look at one real data example using the above test. For Carla data, we choose 500 probes and average the expressions based on 2 replicates for each condition, and standardize the expression vector of each probe. Because the dimensionality of the data decreases by one when we subtract the mean vector, we apply an orthogonal transformation to the data with the last column corresponding to $\frac{1}{2}(1, 1, 1, 1)'$, which forces the last entry of the transformed vector to be zero due to subtracting the mean vector. Hence we get 3-dimensional directional data on the unit 3-sphere. Perform the clustering by the size of spherical cap. In this case, it is hard to plot the dendrogram since there are too many objects. So we give the log-log plot of the nodesize vs the linking distance (see Figure 3.4). Each dot in the figure represents a cluster, the x-axis and y-axis represent its corresponding linking distance and nodesize. Here also we apply the same idea to cut the linking distance at $d = 0.0022$

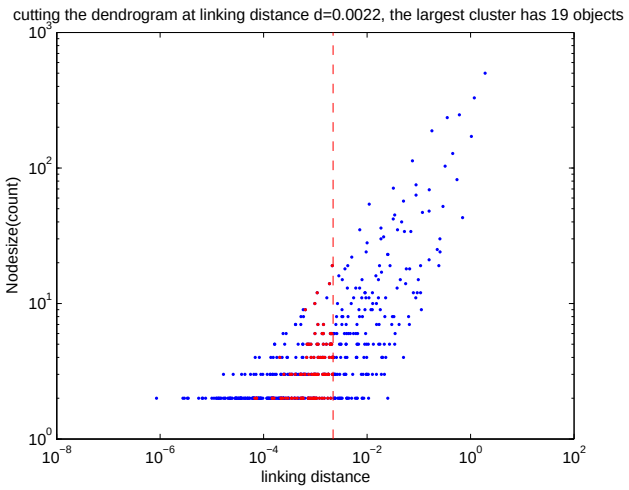

Figure 3.4: Log-Log plot of node size vs linking distance.

which is represented by the vertical red dash line, and the resultant non-simple clusters are represented by red dots. At this level of linking, the largest cluster has 19 points. Applying the above approximation, we get the corresponding p-value less than 10^{-6}, i.e. $P(\Theta_{19} < \theta_{obs}) < 10^{-6}$ using (3.14). So the null hypothesis H_0 is rejected and this means that the directional data is not uniformly distributed on the unit sphere.

Test based on inter-event distance: Sometimes, objects are also called events. In this case, each point on the unit sphere represents an event. When $p = 3$, the theoretical cumulative distribution function (CDF) of inter-event distance is

$$P(\Theta < \theta) = (1 - \cos\theta)/2, \qquad \text{for } 0 \leq \theta \leq \pi. \tag{3.15}$$

In the meantime, we can get the corresponding empirical CDF. Then we can plot the empirical CDF vs theoretical CDF. If the data are really from a homogeneous Poisson point process, we should expect to see a straight diagonal line. For Carla dataset, the plot is displayed by the solid curve in Figure 3.5. We can see systematic deviation from the diagonal dash line. Kolmogorov-Smirnov test for the empirical CDF and theoretical CDF returns a p-value less than 10^{-6}.

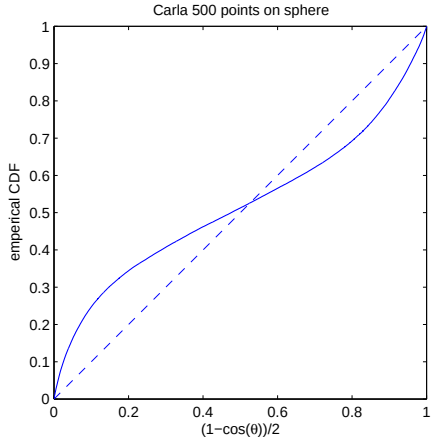

Figure 3.5: Empirical CDF vs theoretical CDF of inter-event distance

3.5 Detection of significant clusters in presence of an inhomogeneous Poisson background

An important problem in clustering of microarray data is that the normalized data is not uniformly distributed on the $p-1$ sphere, as shown in the previous section. This motivates us to develop methods that permit the detection of significant clusters in the presence of a non-uniform background.

Since our focus is one directional data, we use the von Mises-Fisher distribution as a tool to study non-uniform distributions over the sphere.

von Mises-Fisher distribution: A random p-dimensional unit vector $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ (i.e. $\mathbf{x} \in S_p$) is said to have a von Mises-Fisher distribution with mean direction of unit vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)$ and concentration parameter $\kappa > 0$ (denoted as $\mathbf{x} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$) if its density is specified as follows,

$$f_{vMF}(\mathbf{x}; \boldsymbol{\mu}, \kappa) \propto \exp(\kappa \langle \mathbf{x}, \boldsymbol{\mu} \rangle). \tag{3.16}$$

In particular, when $p = 3$, this distribution is named as Fisher distribution $F(\boldsymbol{\mu}, \kappa)$

and its density is

$$f(\mathbf{x}) = \frac{\kappa}{2\sinh\kappa}\exp(\kappa\boldsymbol{\mu}^T\mathbf{x}). \tag{3.17}$$

This density can also be specified in polar coordinates as follows

$$f(\theta,\phi) = \frac{\kappa}{4\pi\sinh\kappa}\exp\left\{\kappa\left[\cos\theta\cos\alpha + \sin\theta\sin\alpha\cos(\phi-\beta)\right]\right\}\sin\theta, \tag{3.18}$$

and denoted as $f_{vMF}([\theta,\phi];[\alpha,\beta],\kappa)$ and $[\theta,\phi] \sim vMF([\alpha,\beta],\kappa)$ by abusing our notation. This can be generalized to high dimensional cases.

Biologists ask questions such as "Is there any cluster with more genes than expected that respond in a similar pattern?". From the previous section, we know that the directional data obtained from preprocessing microarray data does not follow a homogeneous process. So to address biologists' question, we need to detect clusters (with more genes than expected) when the data is from an inhomogeneous Poisson process. In practice, we do not know the underlying density of the inhomogeneous Poisson process and need to estimate this unknown density. In this study, we have considered the following three methods. In the first two methods, the estimated density will be used as the intensity function of the inhomogeneous Poisson process.

Method 1: Model observed data using a mixture of von Mises-Fisher distributions: Empirically, we can visualize several sub-groups of observations in our microarray directional data. This motivates us to postulate that the intensity function of the inhomogeneous Poisson process corresponds to a mixture of von Mises-Fisher distributions, i.e

$$f(\mathbf{x}) = \sum_{j=1}^{K} p_j f_{vMF}(\mathbf{x};\kappa_j,\boldsymbol{\mu}_j), \tag{3.19}$$

where $\sum_{j=1}^{K} p_j = 1$ and K is the number of components. If the fitting succeeds, the estimated density $\hat{f}(\mathbf{x})$ would be used as an inhomogeneous Poisson intensity for computing p-values. Thus we try to fit the data with mixtures of type (3.19).

We resort to WinBUGS to estimate the parameters in the mixture model (3.19). WinBUGS is a Markov Chain Monte Carlo (MCMC) based Bayesian analysis package, in which a prior distribution is needed for each parameter. Here we use non-informative priors on each parameter: uniform prior on each concentration parameter κ_j, uniform prior on each center direction parameter $\boldsymbol{\mu}_j$, and Dirichlet prior on p_j's. More importantly, we need to specify the number of components in the mixture model. Through simulations, we find that this WinBUGS estimation scheme is very sensitive to this specification. More explicitly, when the data is really generated from a mixture of several von Mises-Fisher components and you specify the correct number of components in the mixture, the estimated parameters are very close to the true parameters. But the estimate can quickly deteriorate when the data is contaminated by some additional data points from another unknown component. This indicates that we may include the number of components as an unknown parameters to be estimated. But this involves heavy computation and is not addressed in our work. See Richardson and Green (1997) for discussion on reversible jump MCMC.

To provide better understanding of the sensitivity in the WinBUGS parameter estimation of a mixture of von Mises-Fisher distributions, we do the following experiments:

(1) Our data consists of 200 observations from $vMF((0,-1,0),10)$ and 100 observations form $vMF((0,1,0),20)$. The number of vMF components in WinBUGS codes is set to be 2, which is correct according to our simulation.

(2) We add 30 observations from $vMF((1,0,0),15)$ to the data set in (1), but still set the number of vMF components in the WinBUGS code to be 2, which is wrong in this case.

(3) We replace the 30 observations in (2) by 30 observations from $vMF((1,0,0),0)$ (uniform distribution over the sphere) and set the number of components in the mixture to be 2 in WinBUGS.

	p_1	κ_1	$\boldsymbol{\mu}_1$	p_2	κ_2	$\boldsymbol{\mu}_2$
(1)	0.667	10.621	(-0.015, -0.999, 0.035)	0.333	22.939	(-0.025, 0.999, 0.0234)
(2)	0.696	4.933	(0.138, -0.989, 0.050)	0.304	22.409	(-0.023, 0.999, 0.023)
(3)	0.682	5.959	(-0.030, -0.999, 0.038)	0.318	17.585	(-0.047, 0.999, -0.019)

Table 3.1: WinBUGS estimator of mixture model

After discarding the first 5000 burn-in simulations from the posterior distribution, we use the means of the next 5000 observations to estimate the parameters and they are reported in Table 3.1. From this table, we can see that when we correctly specify the number of components in the mixture model, the means μ_1, μ_2, κ_1 and κ_2 are very close to the true values giving a very good estimation to the mixture model as in (1). On the other hand, WinBUGS estimation scheme does very poorly when you mis-specify the number of components. In (2) and (3), the concentration parameter κ_1 of the first vMF component (the one with smaller concentration parameter κ) deviates a lot from the corresponding true parameter (4.933 and 5.959 compared to 10). One interpretation of this is given as follows. Denote the spherical cap with highest density and containing probability 0.9 under the von Mises-Fisher distribution with the correct $\kappa = 10$ by $A_{p=0.9}$. The probability of $A_{p=0.9}$ under the von Mises-Fisher distribution with the estimated $\kappa = 4.933$ is approximately $\frac{4.933}{10} \times 0.95$, which is considered to have too much error.

Method 2: Estimate $f(\mathbf{x})$ by non-parametric method:

Because of the sensitivity in method 1, we utilize another method to estimate $f(\mathbf{x})$, namely non-parametric estimation.

Suppose we observe a sample of size n: $\{\mathbf{x}_i\}_{i=1}^n$, with each $\mathbf{x}_i$ is directional data in R^3. Denote $(\alpha_i, \beta_i) = \mathbf{u}^{-1}(\mathbf{x}_i)$, the spherical representation of this i-th observation. Then the non-parametric density estimator at point $(\theta, \phi) = \mathbf{u}^{-1}(\mathbf{x})$ is given as follows

$$\hat{f}([\theta, \phi]) = \frac{1}{n} \sum_{i=1}^{n} \frac{\kappa}{4\pi \sinh \kappa} \exp\{\kappa[\cos(\theta)\cos\alpha_i + \sin\theta \sin\alpha_i \cos(\phi - \beta_i)]\} \sin\theta, \quad (3.20)$$

where κ is a smoothing concentration parameter. Note that it can be generalized to use

any other kernel on the unit hyper-sphere. In contrast to the usual smoothing parameter, the estimated non-parametric density is smoother when κ is smaller.

The above non-parametric estimation scheme gives us an estimate of the underlying density of the inhomogeneous Poisson process on the unit sphere, and this permits us to compute p-values for clusters. Notice that we can represent any cluster by a spherical cap (say with center $\mathbf{x}_0$ and central angle θ_0) and the number of observations k_0 it includes. The probability of this spherical cap under the estimated density is given by

$$p_0 = \int_0^{\pi} \int_0^{2\pi} \hat{f}((\theta, \phi)) I[d_{cos}(\mathbf{x}_0, \mathbf{u}((\theta, \phi))) < 1 - \cos\theta_0] d\phi d\theta, \tag{3.21}$$

where $I(\cdot)$ is an indicator function.

Hence the probability of observing this cluster in an inhomogeneous Poisson process driven by density $\hat{f}$ can be approximated by the probability that a Poisson random variable with mean $\lambda_0 = n \cdot p_0$ is larger than or equal to k_0, i.e.

$$P[\#\text{ observations in cap } \geq k_0] = \sum_{i=k_0}^{\infty} \frac{\exp(-\lambda_0)\lambda_0^i}{i!}. \tag{3.22}$$

When we are assessing the significance of a "found" cluster, this probability is interpreted as a p-value.

Method 3: Local test based on area proportions

The above two methods attempt to estimate the density of the background coupled with a Poisson approximation. Our third method is a local one. Given any cluster represented by a spherical cap having center $\mathbf{x}_0$, central angle θ_0 and number of observations k_0, we can easily compute A_0, the surface area of the spherical cap. Next we can check another spherical cap with the same center but a larger surface area rA_0 ($r > 1$) and assume that it includes k_r observations. Are the frequencies in the two spherical caps consistent with area ratios? In this way, we can assess the significance of this cluster by a

binomial approximation. Namely, imagine that we throw k_r points uniformly in the larger spherical cap. This is a binomial trial with k_r repetitions and the probability of success being $\frac{1}{r}$. The significance can be assessed by the probability that there are at least k_0 points falling inside the smaller spherical cap and this probability is

$$P[\#\text{observations in cap } \geq k_0] = \sum_{j=k_0}^{k_r} \begin{pmatrix} k_r \\ j \end{pmatrix} (\frac{1}{r})^k (1 - \frac{1}{r})^{k_r - j}. \tag{3.23}$$

3.6 Simulation

We simulate a data set consisting of 550 observations: 500 from $vMF([\pi/2, \pi], 10)$ and 50 from $vMF([\pi/2, (1-\gamma)\pi], 100)$, where γ is a parameter controlling the distance between the centers of these two vMF's. The contour plot of the true densities is given in Figure 3.6. In this simulation, we treat the big von Mises-Fisher component as the "background", and we want to detect the more concentrated but smaller von Mises-Fisher component in the presence of the big one, and also diminish the significance of the clusters produced from the background. Here we investigate four different cases: $\gamma = 1, 0.5, 0.25$, or 0.125. As γ gets smaller, the two components in the mixture model moves closer to each other and it becomes harder to distinguish . We test both method 2 with smoothing parameter $\kappa = 5, 10, 20, 30$, or 40 and method 3 with area ratio $r = 2, 4, 8$, or 16.

For each case, we cut the dendrogram at linking distance $d = 2^{-5}$. We only check the resulting distinct clusters after cutting and excluding the clusters with less than 8 observations inside. Of these clusters, we divide them into two types: type 1 means that all the observations in this cluster belong to the "background" (big von Mises-Fisher component), and type 2 means that this cluster contains at least one observation from the small von Mises-Fisher component. Note that we are interested in reducing the significance (detection) of clusters of type 1 while continuing to detect clusters of type 2.

We observe that combining these two methods does a good job of meeting this objec-

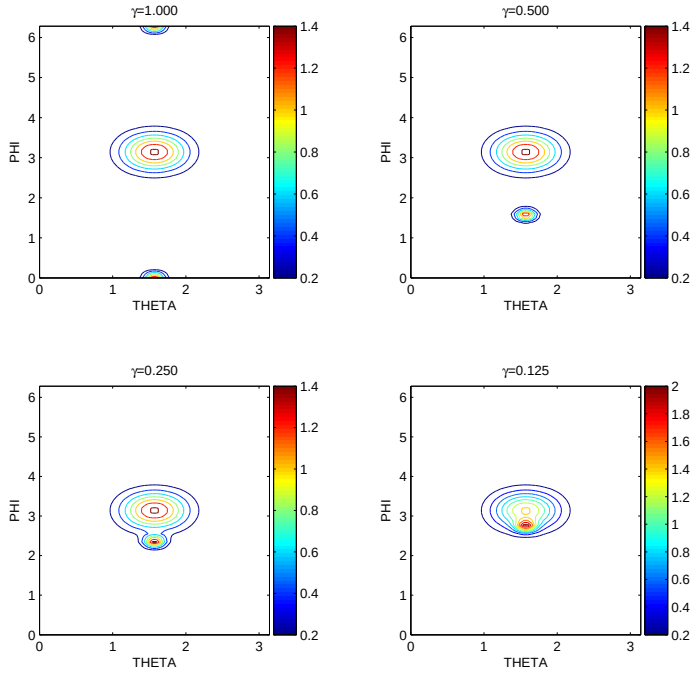

Figure 3.6: True density of simulation

tive. Here a cluster in the dendrogram is called significant when the corrected p-values using both smoothing and local test are smaller than a pre-specified threshold value, for which we use 0.01.

In particular, for $\gamma = 1$, we get the best result with smoothing parameter $\kappa = 20$ and area ratio $r = 4$. In this case, after cutting the dendrogram and excluding the clusters with less than 8 observations, there are 16 distinct clusters (14 type-1 clusters and 2 type-2 clusters). Out of the 14 type-1 clusters, the smoothing method with $\kappa = 20$ flags 7 significant and 7 non-significant; the local test with area ratio $r = 4$ flags 3 significant and 11 non-significant. Intersect these two tests together, we only flag 1 of the 14 type-1 clusters significant. For the 2 type-2 clusters, both smoothing and local test methods call them significant. This is reported in Table 3.2.

For the case that $\gamma = 0.5$, we have exactly the same result as the case $\gamma = 1$. When κ becomes smaller, the job becomes much harder. The best parameters for $\gamma = 0.25$ are

	$\gamma = 1$		
	total clusters	significant	not significant
Cluster type 1	14	1 (7, 3)	13 (7, 11)
Cluster type 2	2	2 (2, 2)	0 (0, 0)
	$\gamma = 0.5$		
Cluster type 1	14	1 (7, 3)	13 (7, 11)
Cluster type 2	2	2 (2, 2)	0 (0, 0)
	$\gamma = 0.25$		
Cluster type 1	13	1 (8,1)	12 (5, 12)
Cluster type 2	3	2 (3,2)	1 (0, 1)
	$\gamma = 0.125$		
Cluster type 1	12	3 (9,3)	9 (3, 9)
Cluster type 2	4	3 (3,3)	1 (1, 1)

Table 3.2: Detection result of different cases. Here cluster type 1 means a cluster consisting of observations only from the big vMF component; cluster type 2 means a cluster at least containing some observations from the small vMF component. Each cell $a(b, c)$ means that the numbers of clusters called significant (or not significant) are a using both method, b for smoothing, and c for local test repectively.

$\kappa = 5$ or 10 and $r = 2$; for case $\gamma = 0.125$ the parameters $\kappa = 5$ and $r = 8$ give the best result. See Table 3.2 for their corresponding performance.

We now take a closer look at the case with $\gamma = 0.5$ with best parameter κ and r. In Figure 3.7, the top left panel gives the dot plot of each cluster with more than 8 observations, where red means that this cluster is called significant using the *smoothing test*. The corresponding result using the *local test* and *both methods* are shown in the top right and the middle right panels respectively. Since this is based on simulation, we know the detailed information of each cluster: in the middle left panel, the cluster represented by a blue dot means that all the observations in this cluster are from the "background" big vMF; red means all observations are from the small vMF; magenta means this cluster is a mixture of observations from both the big vMF and the small vMF. After cutting the dendrogram at linking distance $d = 2^{-5}$ and excluding the clusters with less than 8 observations, there are only 16 distinct clusters left and they are plotted in the left bottom panel in which magenta means the cluster is of type 2. For these 16 distinct clusters, their

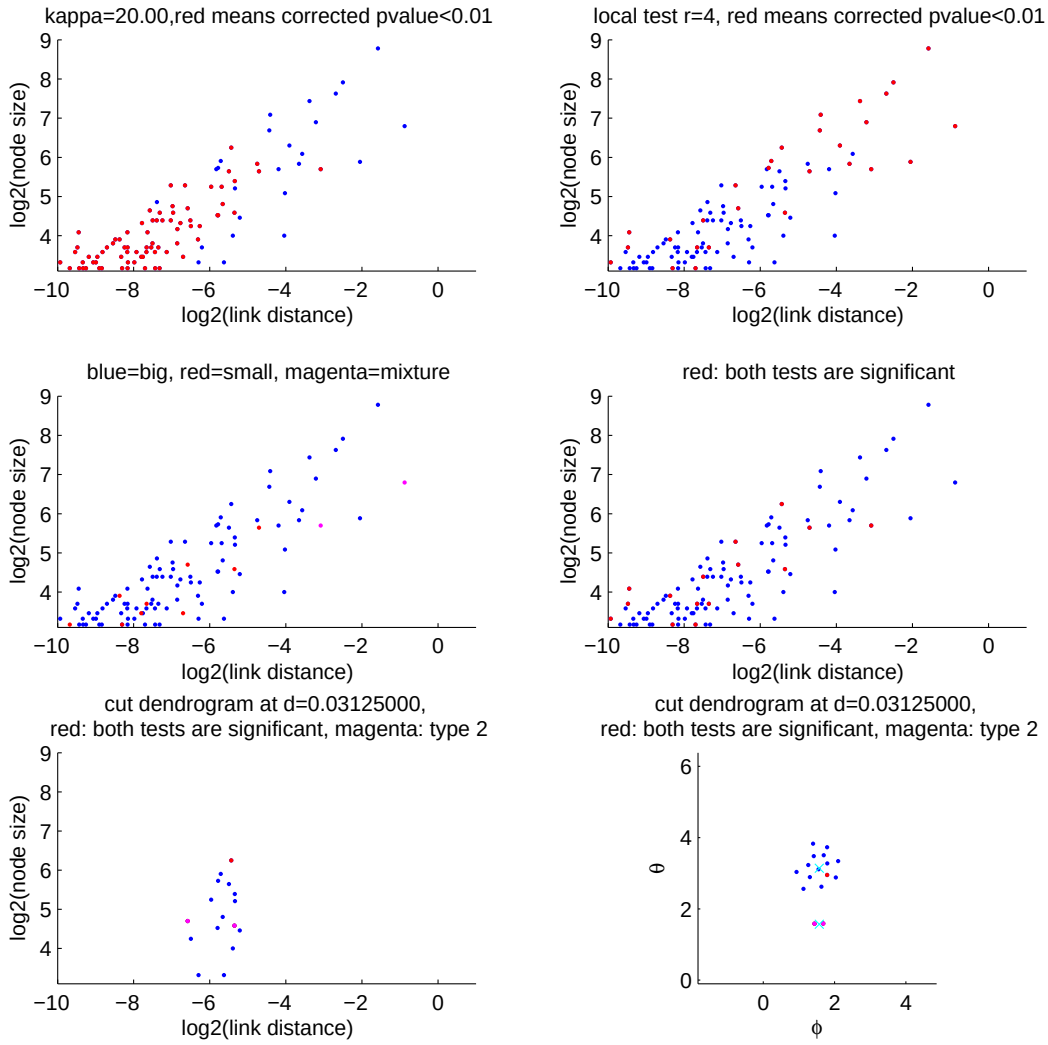

Figure 3.7: Significance test result for $\gamma = 0.5$

centers in polar coordinates are plotted in the bottom right panel with magenta meaning type 2. The two cyan crosses denote the centers of the two von Mises-Fisher distributions.

3.7 Summary

In this chapter, motivated by studying microarray data, we obtain directional data on the unit sphere after pre-processing microarray data and propose the idea of linking by the size of spherical cap to study clustering of directional data. In this way, the directional data is on the unit sphere and we can define the probability of each cluster by assuming a density over the unit sphere. Empirically we observe that the directional data obtained from preprocessing microarray data is not uniformly distributed on the unit sphere. This implies that the data follows an inhomogeneous Poisson process driven by some unknown density and it poses a lot of difficulty for us. To estimate this unknown density, we consider two methods: the mixture model and non-parametric estimation. According to our simulation results, we find that combining the non-parametric method with a local test helps us to reduce the significance of background clusters while detecting more concentrated ones.

Chapter 4

Summary of Dissertation

In chapter 2, we study the stochastic volatility model in mathematical finance. We have obtained an upper bound for polynomial convergence rates of the proposed Gaussian approximation scheme. The argument we adopted limits the sharpness of the derived bound. More specifically, the discrete sums used to approximate the option price integrals involve a two-step Markov chain whose spectral gap is zero. Hence the preferred method of spectral analysis for the Markov transition probability operator does not apply. Instead, we resort to the "big-block-small-block" splitting technique to deal with functionals of the underlying strong mixing process.

In chapter 3, we study clustering directional data on the unit sphere. We propose to link by the size of spherical cap in clustering directional data and to study the probability of each cluster represented by a spherical cap. Empirically, the directional data obtained from preprocessing microarray data is not uniformly distributed on the unit sphere. To overcome this difficulty, we propose a method of combining a non-parametric estimation with a local test to reduce the significance of background clusters while detecting more concentrated ones.

Appendix A

Proofs of Theorems and Lemmas

Proof of **Lemma 1** : Let $\mu_0 = \frac{a}{1-b}$, $\sigma_0^2 = \frac{c^2}{1-b^2}$; and for $n = 1, 2, ...$, let $\mu_n = a\frac{1-b^n}{1-b} + b^n h_0$, $\sigma_n^2 = c^2\frac{1-b^{2n}}{1-b^2}$, $\chi_{1n} = b^n(h_0 - \frac{a}{1-b})$ and $\chi_{2n} = -\frac{c^2 b^{2n}}{1-b^2}$. Then $\mu_n = \mu_0 + \chi_{1n}$, $\sigma_n^2 = \sigma_0^2 + \chi_{2n}$, and $h_n|h_0 \sim N(\mu_n, \sigma_n^2)$. It follows from the Markov property of h that

$$
\begin{aligned}
& \alpha(n) \\
= & \sup_{A\in\sigma(h_0), B\in\sigma(h_n)} |P(A\cap B) - P(A)P(B)| \\
= & \sup_{A\in\sigma(h_0), B\in\sigma(h_n)} \left| \int_A \int_B f_{h_0}(x) f_{h_n|h_0}(y|x) dy dx - \int_A f_{h_0}(x) dx \int_B f_{h_n}(y) dy \right| \\
\leq & \sup_{A\in\sigma(h_0), B\in\sigma(h_n)} \int_A f_{h_0}(x) \int_B \left| f_{h_n|h_0}(y|x) - f_{h_n}(y) \right| dy dx \\
\leq & \int_{-\infty}^{\infty} f_{h_0}(x) \int_{-\infty}^{\infty} \left| f_{h_n|h_0}(y|x) - f_{h_n}(y) \right| dy dx \\
\leq & \int_{-\infty}^{\infty} f(x;\mu_0,\sigma_0^2) \int_{-\infty}^{\infty} \left| f(y;\mu_n,\sigma_n^2) - f(y;\mu_0,\sigma_0^2) \right| dy dx \\
= & \int_{-\infty}^{\infty} f(x;\mu_0,\sigma_0^2) \int_{-\infty}^{\infty} \left| \int_0^1 \left[\left.\frac{\partial f}{\partial \mu}\right|_{(y,\mu_0+t\chi_{1n},\sigma_0^2+t\chi_{2n})} \cdot \chi_{1n} \right.\right. \\
& \left.\left. + \left.\frac{\partial f}{\partial(\sigma^2)}\right|_{(y,\mu_0+t\chi_{1n},\sigma_0^2+t\chi_{2n})} \cdot \chi_{2n} \right] dt \right| dy dx \\
\leq & \int_{-\infty}^{\infty} f(x;\mu_0,\sigma_0^2) \int_{-\infty}^{\infty} \int_0^1 \left(\left| \left.\frac{\partial f}{\partial \mu}\right|_{(y,\mu_0+t\chi_{1n},\sigma_0^2+t\chi_{2n})} \right| \cdot |\chi_{1n}| \right. \\
& \left. + \left| \left.\frac{\partial f}{\partial(\sigma^2)}\right|_{(y,\mu_0+t\chi_{1n},\sigma_0^2+t\chi_{2n})} \right| \cdot |\chi_{2n}| \right) dt dy dx
\end{aligned}
$$

$$
\begin{aligned}
&= |b|^n \int_{-\infty}^{\infty} f(x;\mu_0,\sigma_0^2) \int_{-\infty}^{\infty}\int_0^1 \left| \left.\frac{\partial f}{\partial \mu}\right|_{(y,\mu_0+t\chi_{1n},\sigma_0^2+t\chi_{2n})} \right| \cdot \left| x - \frac{a}{1-b} \right| \, dtdydx \\
&+ |b|^{2n} \int_{-\infty}^{\infty} f(x;\mu_0,\sigma_0^2) \int_{-\infty}^{\infty}\int_0^1 \left| \left.\frac{\partial f}{\partial (\sigma^2)}\right|_{(y,\mu_0+t\chi_{1n},\sigma_0^2+t\chi_{2n})} \right| \cdot \left| \frac{-c^2}{1-b^2} \right| \, dtdydx \\
&= |b|^n c_2 + |b|^{2n} c_3 \qquad \text{for some } c_2 > 0 \text{ and } c_3 > 0 \\
&\leq c_1 |b|^n \qquad \text{for some } c_1 > 0. \qquad \square
\end{aligned}
$$

Proof of **Lemma 2** : Direct calculation as in Lee, Cheng, and Ji (2004) immediately proves (2.29) — (2.33). (2.34) follows from that for sufficiently large j, $\left|\exp(\frac{c^2 b^j}{1-b^2}) - 1\right| \leq 2\frac{c^2|b|^j}{1-b^2}$. In fact, we can obtain asymptotic upper bounds for the absolute values of differences between those moments and their limits respectively. For instance,

$$
|\text{corr}(U_n, V_n) - \rho_0| \leq c_1/n, \tag{A.1}
$$

for some $c_1 > 0$. $\square$

Proof of **Lemma 3** : For $\epsilon \in (0, 1/2)$, let

$$
p = p_n = [n^{\frac{1}{2}+\epsilon}], \quad q = q_n = [n^{\frac{1}{2}-\epsilon}], \quad k = k_n = \left[\frac{n}{p+q}\right],
$$

where $[x]$ denotes the integer part of $x \in I\!R$. Note that $k_n = n^{\frac{1}{2}-\epsilon}(1+o(1))$ as $n \to \infty$. Without loss of generality, we can assume that $n = p + q \pmod 0$, hence $k = \frac{n}{p+q}$. For $m = 0, 1, ..., k-1$, define big blocks

$$
\begin{aligned}
\xi_m^U &= \sum_{j=m(p+q)}^{(m+1)p+mq-1} x_j' \\
\xi_m^V &= \sum_{j=m(p+q)}^{(m+1)p+mq-1} x_j''
\end{aligned}
$$

and small blocks

$$\begin{aligned}\zeta_m^U &= \sum_{j=(m+1)p+mq}^{(m+1)(p+q)-1} x_j^{'} \\ \zeta_m^V &= \sum_{j=(m+1)p+mq}^{(m+1)(p+q)-1} x_j^{''}.\end{aligned}$$

Then

$$\begin{aligned}U_n^* &= \sigma_{U_n}^{-1}(\xi_0^U + \xi_1^U + \cdots + \xi_{k-1}^U) + \sigma_{U_n}^{-1}(\zeta_0^U + \zeta_1^U + \cdots + \zeta_{k-1}^U) \triangleq U_n^{'} + U_n^{''}, \\ V_n^* &= \sigma_{V_n}^{-1}(\xi_0^V + \xi_1^V + \cdots + \xi_{k-1}^V) + \sigma_{V_n}^{-1}(\zeta_0^V + \zeta_1^V + \cdots + \zeta_{k-1}^V) \triangleq V_n^{'} + V_n^{''}.\end{aligned}$$

Let $\phi(u,v) = \exp[-\frac{1}{2}(u^2+2\rho_0 uv+v^2)]$ be the characteristic function of bivariate normal distribution with mean $(0,0)$ and covariance matrix $\begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}$. Triangle inequalities imply that

$$|E\exp[i\ (uU_n^* + vV_n^*)] - \phi(u,v)| \leq \mathrm{I}\ +\ \mathrm{II}\ + \mathrm{III} \tag{A.2}$$

where

$$\begin{aligned}\mathrm{I} &= \left|E\exp[i\ (uU_n^* + vV_n^*)] - E\exp[i\ (uU_n^{'} + vV_n^{'})]\right|, \\ \mathrm{II} &= \left|E\exp[i\ (uU_n^{'} + vV_n^{'})] - \prod_{m=0}^{k-1} E\exp\left[i\ \left(u\frac{\xi_m^U}{\sigma_{U_n}} + v\frac{\xi_m^V}{\sigma_{V_n}}\right)\right]\right|, \\ \mathrm{III} &= \left|\prod_{m=0}^{k-1} E\exp\left[i\ \left(u\frac{\xi_m^U}{\sigma_{U_n}} + v\frac{\xi_m^V}{\sigma_{V_n}}\right)\right] - \phi(u,v)\right|.\end{aligned}$$

We will obtain an upper bound for each part, in the order of parts II, III and I, respectively.

Setting $\prod_{j=m}^{m-1} a_j = 1$ and $\sum_{j=m}^{m-1} a_j = 0$ for any $m = 0, 1, ..., k-1$, we have

$$
\begin{aligned}
\mathrm{II} &= \left| \sum_{m=0}^{k-1} \left[\left(\prod_{j=0}^{m-1} E \exp\left(i \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right) E \exp\left(i \sum_{j=m}^{k-1} \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right. \right. \\
&\quad \left. \left. - \left(\prod_{j=0}^{m} E \exp\left(i \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right) E \exp\left(i \sum_{j=m+1}^{k-1} \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right] \right| \\
&\leq \sum_{m=0}^{k-1} \left| \left(\prod_{j=0}^{m-1} E \exp\left(i \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right) E \exp\left(i \sum_{j=m}^{k-1} \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right. \\
&\quad \left. - \left(\prod_{j=0}^{m} E \exp\left(i \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right) E \exp\left(i \sum_{j=m+1}^{k-1} \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right| \\
&\leq \sum_{m=0}^{k-1} \left| E \exp\left(i \sum_{j=m}^{k-1} \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right. \\
&\quad \left. - E \exp\left(i \left(u \frac{\xi_m^U}{\sigma_{U_n}} + v \frac{\xi_m^V}{\sigma_{V_n}} \right) \right) E \exp\left(i \sum_{j=m+1}^{k-1} \left(u \frac{\xi_j^U}{\sigma_{U_n}} + v \frac{\xi_j^V}{\sigma_{V_n}} \right) \right) \right| \\
&\leq k\alpha(q) \leq C_{\mathrm{II}}\, n^{\frac{1}{2}-\epsilon}\, \exp\left(-b_1 n^{\frac{1}{2}-\epsilon} \right) \qquad \text{(A.3)}
\end{aligned}
$$

for some constant $C_{\mathrm{II}} > 0$.

For part III, we need to evaluate several moments asymptotically. First, it follows from (2.31)–(2.36) that

$$
\begin{aligned}
E\left[(\xi_0^U)^2\right] &= \sigma_U^2\ p\ (1+o(1)) = \sigma_U^2\ n^{\frac{1}{2}+\epsilon}\ (1+o(1)), \\
E\left[(\xi_0^V)^2\right] &= \sigma_V^2\ p\ (1+o(1)) = \sigma_V^2\ n^{\frac{1}{2}+\epsilon}\ (1+o(1)), \\
E[\xi_0^U \xi_0^V] &= \rho_0\ \sigma_U \sigma_V\ p\ (1+o(1)) = \rho_0\ \sigma_U \sigma_V\ n^{\frac{1}{2}+\epsilon}\ (1+o(1)).
\end{aligned}
$$

Next,

$$
\begin{aligned}
|E\left[(\xi_0^U)^3\right]| &= \left| E\left[\left(\sum_{j=0}^{p-1} x_j' \right)^3 \right] \right| = \left| \sum_{0 \leq m,j,k \leq p-1} E(x_m' x_j' x_k') \right| \\
&\leq 6 \sum_{0 \leq m \leq j \leq k \leq p-1} \left| E(x_m' x_j' x_k') \right|.
\end{aligned}
$$

For any $0 < l < p$, there are no more than $l^2 p$ terms for which $\max\{j-m, k-j\} \leq l$. The remaining terms can be represented in the form $\left|Ex'_m x'_j x'_k\right| = |E\vartheta\varsigma|$, where either $\vartheta = x'_m$, $\varsigma = x'_j x'_k$, or $\vartheta = x'_m x'_j$, $\varsigma = x'_k$, depending on whether $j - m > l$ or $k - j > l$. Following Lemma 1.3 in Ibragimov (1962), we have

$$\begin{aligned} |E(\vartheta\varsigma)| &\leq |E\vartheta|\,|E\varsigma| + c_0\ (\alpha(l))^{1/2} \\ &= c_0\ (\alpha(l))^{1/2} \end{aligned}$$

for some $c_0 > 0$. Setting $l = O(p^{\delta_2})$ for some $\delta_2 \in (0,1)$ yields

$$E\left[(\xi_0^U)^3\right] = O((p^{\delta_2})^2 p) = O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)}).$$

Similar arguments work for $E\left[(\xi_0^U)^2 \xi_0^V\right]$, $E\left[\xi_0^U (\xi_0^V)^2\right]$ and $E\left[(\xi_0^V)^3\right]$.

Now the following Taylor expansions are in order.

$$\begin{aligned} \psi_{1n}(u,v) &\triangleq E\exp\left[i\ \left(u\frac{\xi_0^U}{\sigma_{U_n}} + v\frac{\xi_0^V}{\sigma_{V_n}}\right)\right] \\ &= 1 - \frac{1}{2}\left[\frac{u^2}{\sigma_{U_n}^2}E(\xi_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\xi_0^U\xi_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\xi_0^V)^2\right] \\ &\quad + \frac{\Theta_1}{6}\left[\frac{u^3}{\sigma_{U_n}^3}E(\xi_0^U)^3 + \ 3\frac{u^2 v}{\sigma_{U_n}^2\sigma_{V_n}}E(\xi_0^U)^2\xi_0^V\right. \\ &\quad \left. + \ 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\xi_0^U(\xi_0^V)^2 + \ \frac{v^3}{\sigma_{V_n}^3}E(\xi_0^V)^3\right] \quad \text{where} \quad |\Theta_1| \leq 1 \\ &= 1 + O(n^{2\varrho+\frac{1}{2}+\epsilon-1}) + O(n^{3\varrho+(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}}). \end{aligned} \tag{A.4}$$

Apparently we need to impose the constraints

$$\begin{aligned} 2\varrho + 1/2 + \epsilon - 1 &< 0, \\ 3\varrho + (1+2\delta_2)(1/2+\epsilon) - 3/2 &< 0. \end{aligned}$$

Therefore,

$$
\begin{aligned}
& \log \psi_{1n}(u,v) \\
= \; & -\frac{1}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\xi_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\xi_0^U\xi_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\xi_0^V)^2\right) \\
& +\frac{\Theta_1}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\xi_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\xi_0^U)^2\xi_0^V + 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\xi_0^U(\xi_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\xi_0^V)^3\right) \\
& +\Theta_2\left[-\frac{1}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\xi_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\xi_0^U\xi_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\xi_0^V)^2\right)\right. \\
& + \frac{\Theta_1}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\xi_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\xi_0^U)^2\xi_0^V\right. \\
& \left.\left.+3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\xi_0^U(\xi_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\xi_0^V)^3\right)\right]^2 \quad (\text{where } |\Theta_2| \le 1) \\
= \; & -\frac{1}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\xi_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\xi_0^U\xi_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\xi_0^V)^2\right) \\
& +\frac{\Theta_1}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\xi_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\xi_0^U)^2\xi_0^V + 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\xi_0^U(\xi_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\xi_0^V)^3\right) \\
& + \text{higher order term}, \qquad (A.5)
\end{aligned}
$$

which implies that

$$
\begin{aligned}
& \log\left\{\prod_{m=0}^{k-1} E\exp\left[i\ \left(u\frac{\xi_m^U}{\sigma_{U_n}} + v\frac{\xi_m^V}{\sigma_{V_n}}\right)\right]\right\} - \ \frac{-1}{2}(u^2 + 2\rho_0 uv + v^2) \\
= \; & k\log\psi_{1n}(u,v) + \frac{1}{2}(u^2 + 2\rho_0 uv + v^2) \\
= \; & \frac{1}{2}(u^2 + 2\rho_0 uv + v^2) - \frac{k}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\xi_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\xi_0^U\xi_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\xi_0^V)^2\right) \\
& +\frac{k\Theta_1}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\xi_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\xi_0^U)^2\xi_0^V + 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\xi_0^U(\xi_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\xi_0^V)^3\right) \\
& + \text{higher order term} \\
= \; & \frac{1}{2}\left[u^2\left(1 - \frac{kE(\xi_0^U)^2}{\sigma_{U_n}^2}\right) + 2uv\left(\rho_0 - \frac{kE\xi_0^U\xi_0^V}{\sigma_{U_n}\sigma_{V_n}}\right) + v^2\left(1 - \frac{kE(\xi_0^V)^2}{\sigma_{V_n}^2}\right)\right] \\
& +\frac{k\Theta_1}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\xi_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\xi_0^U)^2\xi_0^V + 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\xi_0^U(\xi_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\xi_0^V)^3\right) \\
& + \text{higher order term} \qquad (A.6)
\end{aligned}
$$

Asymptotic estimates of $1 - \frac{kE(\xi_0^U)^2}{\sigma_{U_n}^2}$, $\rho_0 - \frac{kE\xi_0^U\xi_0^V}{\sigma_{U_n}\sigma_{V_n}}$ and $1 - \frac{kE(\xi_0^V)^2}{\sigma_{V_n}^2}$ are needed for the completion of (A.6).

First note that

$$\begin{aligned}
& \left|\frac{E(\xi_0^U + \cdots + \xi_{k-1}^U)^2}{\sigma_{U_n}^2} - 1\right| \\
= & \left|\frac{E(\xi_0^U + \cdots + \xi_{k-1}^U)^2}{\sigma_{U_n}^2} - VarU_n^*\right| \\
= & \left|Var(U_n^{''}) - 2\ E(U_n^* U_n^{''})\right| \\
\leq & \ Var(U_n^{''}) + 2\sqrt{E(U_n^*)^2\ E(U_n^{''})^2} \\
= & \ Var(U_n^{''}) + 2\sqrt{Var(U_n^{''})} \\
= & \ O(n^{-\epsilon}), \qquad \text{(A.7)}
\end{aligned}$$

since equation (1.5) in Ibragimov (1962) and (2.31) imply that

$$Var(U_n^{''}) = O(k^2\ (\alpha(p))^{1/2}) + O\left(\frac{kn^{1/2-\epsilon}}{n}\right) = O(n^{-2\epsilon}).$$

Next, equation (1.5) in Ibragimov (1962) implies

$$E\xi_0^U\xi_m^U = O((\alpha(mq + (m-1)p))^{1/2)}).$$

Since

$$E(\xi_0^U + \cdots + \xi_{k-1}^U)^2 = k\ E(\xi_0^U)^2 + 2\sum_{j=1}^{k-1}(k-j)\ E\xi_0^U\xi_j^U,$$

we have

$$\begin{aligned}
\frac{E(\xi_0^U + \cdots + \xi_{k-1}^U)^2}{kE(\xi_0^U)^2} &= 1 + O\left(\sum_{j=1}^{k-1}(1 - j/k)(\alpha(jq + (j-1)p))^{1/2}\right) \\
&= 1 + O(\exp(-n^{1/2-\epsilon})). \qquad \text{(A.8)}
\end{aligned}$$

Putting (A.7) and (A.8) together yields

$$\left| \frac{k\ E(\xi_0^U)^2}{\sigma_{U_n}^2} - 1 \right| = O(n^{-\epsilon}).$$

The same argument yields

$$\left| \frac{kE(\xi_0^V)^2}{\sigma_{V_n}^2} - 1 \right| = O(n^{-\epsilon}).$$

A similar method leads to

$$\frac{E(\xi_0^U + \cdots + \xi_{k-1}^U)(\xi_0^V + \cdots + \xi_{k-1}^V)}{kE(\xi_0^U)(\xi_0^V)} = 1 + O(\exp(-n^{1/2-\epsilon})).$$

For $\rho_0 - \frac{k\ E\xi_0^U\xi_0^V}{\sigma_{U_n}\sigma_{V_n}}$, we recall (A.1) and evaluate

$$\frac{E(\xi_0^U + \cdots + \xi_{k-1}^U)(\xi_0^V + \cdots + \xi_{k-1}^V)}{E(U_n^c V_n^c)} - 1, \tag{A.9}$$

where (the superscript "c" stands for "centered")

$$\begin{aligned} U_n^c &= \sum_{j=0}^{n-1} x_j' = \sum_{j=0}^{n-1} \left(e^{h_j} - e^{\frac{a}{1-b} + \frac{c^2}{2(1-b^2)}} \right) \\ V_n^c &= \sum_{j=0}^{n-1} x_j'' = \sum_{j=0}^{n-1} \left(e^{h_j/2} \epsilon_{j+1} \right). \end{aligned}$$

The estimation for (A.9) can be obtained as follows:

$$
\begin{aligned}
& \left| 1 - \frac{E(\xi_0^U + \cdots + \xi_{k-1}^U)(\xi_0^V + \cdots + \xi_{k-1}^V)}{EU_n^c V_n^c} \right| \\
= & \left| \frac{EU_n^c V_n^c - E(\xi_0^U + \cdots + \xi_{k-1}^U)(\xi_0^V + \cdots + \xi_{k-1}^V)}{EU_n^c V_n^c} \right| \\
= & \left| \frac{E(\sum_{j=0}^{k-1} \xi_j^U + \sum_{j=0}^{k-1} \zeta_j^U)(\sum_{j=0}^{k-1} \xi_j^V + \sum_{j=0}^{k-1} \zeta_j^V) - E(\sum_{j=0}^{k-1} \xi_0^U)(\sum_{j=0}^{k-1} \xi_0^V)}{EU_n^c V_n^c} \right| \\
= & \left| \frac{E(\sum_{j=0}^{k-1} \xi_j^U)(\sum_{j=0}^{k-1} \zeta_j^V) + E(\sum_{j=0}^{k-1} \zeta_j^U)(\sum_{j=0}^{k-1} \xi_j^V) + E(\sum_{j=0}^{k-1} \zeta_j^U)(\sum_{j=0}^{k-1} \zeta_j^V)}{EU_n^c V_n^c} \right| \\
\leq & \left(\sqrt{E \left(\sum_{j=0}^{k-1} \xi_j^U \right)^2 E \left(\sum_{j=0}^{k-1} \zeta_j^V \right)^2} + \sqrt{E \left(\sum_{j=0}^{k-1} \zeta_j^U \right)^2 E \left(\sum_{j=0}^{k-1} \xi_j^V \right)^2} \right. \\
& \left. + \sqrt{E \left(\sum_{j=0}^{k-1} \zeta_j^U \right)^2 E \left(\sum_{j=0}^{k-1} \zeta_j^V \right)^2} \right) / \left(EU_n^c V_n^c \right) \\
= & \frac{\sqrt{O(n)\ O(n^{1-2\epsilon})} + \sqrt{O(n^{1-2\epsilon})\ O(n)} + \sqrt{O(n^{1-2\epsilon})\ O(n^{1-2\epsilon})}}{O(n)} \\
= & O(n^{-\epsilon}),
\end{aligned}
$$

hence

$$
\left| \frac{k\ E\xi_0^U \xi_0^V}{\sigma_{U_n} \sigma_{V_n}} - \rho_0 \right| = O(n^{-\epsilon}).
$$

Putting all the above rates into (A.6), we get

$$
\begin{aligned}
& \log \prod_{m=0}^{k-1} E \exp \left[i \ \left(u \frac{\xi_m^U}{\sigma_{U_n}} + v \frac{\xi_m^V}{\sigma_{V_n}} \right) \right] - \frac{-1}{2} (u^2 + 2\rho_0 uv + v^2) \\
= & \ (u^2 + v^2)\ O(n^{-\epsilon}) + (|u| + |v|)^3\ O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon}). \qquad \text{(A.10)}
\end{aligned}
$$

Since $|e^s - 1| \leq |s| e^{|s|} \quad \forall s \in I\!R$,

$$
\begin{aligned}
& \left| \prod_{m=0}^{k-1} E \exp \left[i \ \left(u \frac{\xi_m^U}{\sigma_{U_n}} + v \frac{\xi_m^V}{\sigma_{V_n}} \right) \right] \ \exp \left(\frac{1}{2} (u^2 + 2\rho_0 uv + v^2) \right) - 1 \right| \\
= & \ (u^2 + v^2)\ O(n^{-\epsilon}) + (|u| + |v|)^3\ O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon}).
\end{aligned}
$$

Therefore,

$$\begin{aligned}
&\left|\prod_{m=0}^{k-1} E\exp\left[i\ \left(u\frac{\xi_m^U}{\sigma_{U_n}}+v\frac{\xi_m^V}{\sigma_{V_n}}\right)\right]-\exp\left(\frac{-1}{2}(u^2+2\rho_0uv+v^2)\right)\right| \\
= &\ \exp\left(-\frac{1}{2}(u^2+2\rho_0uv+v^2)\right)\ [(u^2+v^2)\ O(n^{-\epsilon}) \\
&+(|u|+|v|)^3\ O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon})]. \qquad \text{(A.11)}
\end{aligned}$$

Finally, we evaluate part I.

$$\begin{aligned}
&\left|Ee^{i(uU_n^*+vV_n^*)}-Ee^{i(uU_n^{'}+vV_n^{'})}\right| \\
= &\ \left|E\left[e^{i(uU_n^{'}+vV_n^{'})}(e^{i(uU_n^{''}+vV_n^{''})}-1)\right]\right| \\
\le &\ \sqrt{E\left|e^{i(uU_n^{'}+vV_n^{'})}\right|^2}\sqrt{E\left|e^{i(uU_n^{''}+vV_n^{''})}-1\right|^2} \\
= &\ \sqrt{\left|E\ (e^{i(uU_n^{''}+vV_n^{''})}-1)(e^{-i(uU_n^{''}+vV_n^{''})}-1)\right|} \\
\le &\ \sqrt{\left|E(e^{i(uU_n^{''}+vV_n^{''})}-1)\right|+\left|E(e^{-i(uU_n^{''}+vV_n^{''})}-1)\right|},
\end{aligned}$$

in which $E(e^{i(uU_n^{''}+vV_n^{''})}-1)$ and $E(e^{-i(uU_n^{''}+vV_n^{''})}-1)$ are essentially the same. It suffices to observe that

$$\begin{aligned}
\left|Ee^{i(uU_n^{''}+vV_n^{''})}-1\right| \ \le &\ \left|Ee^{i(uU_n^{''}+vV_n^{''})}-\prod_{j=0}^{k-1}E\exp\left[i\ \left(u\frac{\zeta_j^U}{\sigma_{U_n}}+v\frac{\zeta_j^V}{\sigma_{V_n}}\right)\right]\right| \\
+ &\ \left|\prod_{j=0}^{k-1}E\exp\left[i\ \left(u\frac{\zeta_j^U}{\sigma_{U_n}}+v\frac{\zeta_j^V}{\sigma_{V_n}}\right)\right]-1\right| \\
\triangleq &\ \mathrm{I}_1\ +\ \mathrm{I}_2.
\end{aligned}$$

Note that the same argument used in dealing with part II will lead to a similar (in fact, sharper) upper bound for I_1. Now we examine I_2.

Let

$$\begin{aligned}\varphi_{1n}(u,v) &= E\exp\left[i\ \left(u\frac{\zeta_0^U}{\sigma_{U_n}}+v\frac{\zeta_0^V}{\sigma_{V_n}}\right)\right],\\ \varphi_n(u,v) &= \prod_{j=0}^{k-1} E\exp\left[i\ \left(u\frac{\zeta_j^U}{\sigma_{U_n}}+v\frac{\zeta_j^V}{\sigma_{V_n}}\right)\right].\end{aligned}$$

Then

$$\begin{aligned}&\varphi_{1n}(u,v)\\ =\ & 1-\frac{1}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\zeta_0^U)^2+2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\zeta_0^U\zeta_0^V+\frac{v^2}{\sigma_{V_n}^2}E(\zeta_0^V)^2\right)\\ &+\frac{\Theta_3}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\zeta_0^U)^3+3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\zeta_0^U)^2\zeta_0^V\right.\\ &\left.+3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\zeta_0^U(\zeta_0^V)^2+\frac{v^3}{\sigma_{V_n}^3}E(\zeta_0^V)^3\right)\end{aligned}$$

with $|\Theta_3|\le 1$. Using (2.31), (2.32) and the argument in part III, we have

$$\begin{aligned}E(\zeta_0^U)^2 &= O(n^{1/2-\epsilon})\\ E\zeta_0^U\zeta_0^V &= O(n^{1/2-\epsilon})\\ E(\zeta_0^V)^2 &= O(n^{1/2-\epsilon})\\ E(\zeta_0^U)^3 &= O(n^{(1+2\delta_3)(1/2-\epsilon)})\\ E(\zeta_0^U)^2\zeta_0^V &= O(n^{(1+2\delta_3)(1/2-\epsilon)})\\ E\zeta_0^U(\zeta_0^V)^2 &= O(n^{(1+2\delta_3)(1/2-\epsilon)})\\ E(\zeta_0^V)^3 &= O(n^{(1+2\delta_3)(1/2-\epsilon)}).\end{aligned}$$

for some $\delta_3\in(0,1)$. Hence

$$\varphi_{1n}(u,v)=1-(u+v)^2\ O(n^{-1/2-\epsilon})+(u+v)^3\ O(n^{(1+2\delta_3)(1/2-\epsilon)-3/2}),$$

$$
\begin{aligned}
& \log \varphi_{1n}(u,v) \\
= \; & \frac{-1}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\zeta_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\zeta_0^U\zeta_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\zeta_0^V)^2\right) \\
& + \frac{\Theta_3}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\zeta_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\zeta_0^U)^2\zeta_0^V\right. \\
& \left. + 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\zeta_0^U(\zeta_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\zeta_0^V)^3\right) + \text{higher order term}
\end{aligned}
$$

and

$$
\begin{aligned}
& \log \varphi_n(u,v) \\
= \; & k \;\log \varphi_{1n}(u,v) \\
= \; & \frac{-k}{2}\left(\frac{u^2}{\sigma_{U_n}^2}E(\zeta_0^U)^2 + 2\frac{uv}{\sigma_{U_n}\sigma_{V_n}}E\zeta_0^U\zeta_0^V + \frac{v^2}{\sigma_{V_n}^2}E(\zeta_0^V)^2\right) \\
& + \frac{k\Theta_3}{6}\left(\frac{u^3}{\sigma_{U_n}^3}E(\zeta_0^U)^3 + 3\frac{u^2v}{\sigma_{U_n}^2\sigma_{V_n}}E(\zeta_0^U)^2\zeta_0^V\right. \\
& \left. + 3\frac{uv^2}{\sigma_{U_n}\sigma_{V_n}^2}E\zeta_0^U(\zeta_0^V)^2 + \frac{v^3}{\sigma_{V_n}^3}E(\zeta_0^V)^3\right) \\
= \; & O((u+v)^2\; n^{1/2-\epsilon+1/2-\epsilon-1}) + O((u+v)^3\; n^{1/2-\epsilon+(1+2\delta_3)(1/2-\epsilon)-3/2}) \\
= \; & O((u+v)^2\; n^{-2\epsilon}) + O((u+v)^3\; n^{\delta_3-2\delta_3\epsilon-2\epsilon-1/2}),
\end{aligned}
$$

which implies that

$$
|\varphi_n(u,v) - 1| = O((u+v)^2 n^{-2\epsilon}) + O((u+v)^3 n^{\delta_3-2\delta_3\epsilon-2\epsilon-1/2}). \tag{A.12}
$$

The proof of Lemma 3 is completed by combining parts I, II and III. □

Proof of **Proposition 1***:* Let Z denote a random vector with distribution $K(\cdot)$ [see (2.40)]. For $\varepsilon > 0$, let K_ε be a smooth kernel probability measure on $\mathbb{R}^2$ such that $K_\varepsilon(B) = K(\varepsilon^{-1}B)$ for every measurable set B. By Corollary 11.5 in Bhattacharya and

Rao (1976), we have

$$\left|\int_{R^2} g\ d(Q_n - \Phi^{(2)})\right| \leq \omega_g(I\!R^2)\ \|\ (Q_n - \Phi^{(2)}) * K_\varepsilon\ \| + 2\ \omega_g^*(2\varepsilon : \Phi^{(2)}) \tag{A.13}$$

for all $\varepsilon > 0$, where $\|\ \|$ is the total variation norm. Choose

$$\varepsilon = 4a'\sqrt{2}n^{-\varrho}. \tag{A.14}$$

Then

$$\hat{K}_\varepsilon(u,v) = 0 \quad \text{if } u^2 + v^2 \geq n^{2\varrho}. \tag{A.15}$$

For every Borel set B and $r \geq 0$, let

$$\begin{aligned} B_1 &= B \cap B(0,r), \\ B_2 &= B \backslash B_1. \end{aligned}$$

Hence

$$\left|(Q_n - \Phi^{(2)})(B_1)\right| \leq \frac{1}{(2\pi)^2}\ |B_1| \int_{R^2} \left|(\hat{Q}_n - \hat{\Phi^{(2)}})(u,v)\ \hat{K}_\varepsilon(u,v)\right|\ dudv \tag{A.16}$$

where $|B_1|$ represents the Lebesgue measure of set B_1. It follows from Lemma 3 that

$$\begin{aligned} & \|\hat{Q}_n - \hat{\Phi^{(2)}}\| \\ = \ & O(k\alpha(q)) \\ + \ & \exp\left(\frac{-1}{2}(u^2 + 2\rho_0 uv + v^2)\right) \\ & \cdot\ [(u^2+v^2)\ O(n^{-\epsilon}) + (|u|+|v|)^3\ O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon})] \\ + \ & \sqrt{O(k\alpha(q)) + O((u+v)^2\ n^{-2\epsilon}) + O((u+v)^3\ n^{\delta_3 - 2\delta_3\epsilon - 2\epsilon - 1/2})}. \end{aligned} \tag{A.17}$$

Therefore,

$$\begin{aligned}
&\int_{R^2} \left|(\hat{Q}_n - \hat{\Phi^{(2)}})(t)\ \hat{K}_\varepsilon(t)\right| dt \\
= &\int_{|t|\le n^\varrho} \left|(\hat{Q}_n - \hat{\Phi^{(2)}})(t)\right| dt \\
= &\ n^{2\varrho}\left[O(k\alpha(q)) + O(n^{-\epsilon}) + O(n^{(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon})\right] \\
&+n^{2\varrho}\ O(\sqrt{k\alpha(q)}) + O(n^{2\varrho}n^{-\epsilon}) + O(n^{3\varrho/2}\ n^{(\delta_3-2\delta_3\epsilon-2\epsilon-1/2)/2}) \\
= &\ O(n^{4\varrho-\epsilon}) + O(n^{2\varrho+(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon}) + O(n^{7\varrho/2+(\delta_3-2\delta_3\epsilon-2\epsilon-1/2)/2}),
\end{aligned}$$

and

$$\begin{aligned}
\left|(Q_n - \Phi^{(2)})(B_1)\right| &\le \frac{1}{(2\pi)^2}\ |B_1| \int_{R^2} \left|(\hat{Q}_n - \hat{\Phi^{(2)}})(t)\ \hat{K}_\varepsilon(t)\right| dt \\
&= O(r^2 n^{4\varrho-\epsilon}) + O(r^2 n^{2\varrho+(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon}) \\
&\quad +O(r^2 n^{7\varrho/2+(\delta_3-2\delta_3\epsilon-2\epsilon-1/2)/2}).
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\left|\left[(Q_n - \Phi^{(2)}) * K_\varepsilon\right](B_2)\right| \\
\le &\ \max\{(Q_n * K_\varepsilon)(B_2),\ (\Phi^{(2)} * K_\varepsilon)(B_2)\} \\
\le &\ \max\left\{P\left((U_n^*)^2 + (V_n^*)^2 \ge \frac{r^2}{4}\right),\ \int_{\{|x|\ge r/2\}} \phi^{(2)}(x)\ dx\right\} \\
&+P\left(|\varepsilon Z| \ge \frac{r}{2}\right),
\end{aligned}$$

where $\phi^{(2)}$ is the density for $\Phi^{(2)}$. By the Berry-Esseen theorem [see Reznik (1968)],

$$\begin{aligned}
P\left((U_n^*)^2 + (V_n^*)^2 \ge \frac{r^2}{4}\right) &\le P\left(|U_n^*| \ge \frac{r}{4}\right) + P\left(|V_n^*| \ge \frac{r}{4}\right) \qquad \text{(A.18)} \\
&\le 2\ \left(C\ n^{-\gamma_2} + \frac{2^{3/2}}{\sqrt{2\pi r^2}} e^{-r^2/18}\right) \int_{\{|x|\ge r/2\}} \phi^{(2)}(x)dx \\
&\le 2^{7/2}(2\pi)^{-1/2} r^{-1} e^{-r^2/16}
\end{aligned}$$

and

$$P\left(|\varepsilon Z| \geq \frac{r}{2}\right) \leq \frac{c_3}{\pi r^3\ n^{3\varrho}} \tag{A.19}$$

for some $c_3 > 0$. Setting $r = 8 \log n$ leads to

$$\left|[(Q_n - \Phi^{(2)}) * K_\varepsilon](B_2)\right| = O(n^{-\gamma_2}) + O(n^{-\varrho}). \tag{A.20}$$

Hence

$$\begin{aligned} & \| (Q_n - \Phi^{(2)}) * K_\varepsilon \| \\ = \ & O(n^{4\varrho-\epsilon}) + O(n^{2\varrho+(1+2\delta_2)(\frac{1}{2}+\epsilon)-\frac{3}{2}+\frac{1}{2}-\epsilon}) \\ & +O(n^{7\varrho/2+(\delta_3-2\delta_3\epsilon-2\epsilon-1/2)/2}) + O(n^{-\gamma_2}) + O(n^{-\varrho}). \qquad \square \end{aligned}$$

Proof of **Lemma 4** : Conditing on $h_0^{(\pi)} = z_2$, define

$$\tau(z_1, z_2) = \inf\{k \geq 1 : \text{sign}(h_k^{(\pi)} - h_k^{(1)}) \neq \text{sign}(z_2 - z_1)\}. \tag{A.21}$$

Then

$$P(\tau(z_1) > n) = \int_{-\infty}^{\infty} P(\tau(z_1, z_2) > n)\ f(z_2; \mu_0, \sigma_0^2)\ dz_2. \tag{A.22}$$

Suppose $z_2 < z_1$, consider the (difference) process $h^{(d)}$ defined by

$$h_k^{(d)} = h_k^{(\pi)} - h_k^{(1)}$$

and let

$$\epsilon_k^{(d)} = \frac{\epsilon_k^{(\pi)} - \epsilon_k^{(1)}}{\sqrt{2}}$$

where $\{\epsilon_k^{(\pi)}\}$ and $\{\epsilon_k^{(1)}\}$ are independent versions of the innovation process $\{\epsilon_k\}$ in (2.27)

associated with $h^{(\pi)}$ and $h^{(1)}$ respectively. Hence

$$h_{k+1}^{(d)} = b\ h_k^{(d)} + \sqrt{2}c\ \epsilon_{k+1}^{(d)} \tag{A.23}$$

with $h_0^{(d)} = z_2 - z_1 < 0$. $\tau(z_1, z_2)$ becomes the first passage time of $h^{(d)}$ at level 0, i.e.

$$\tau(z_1, z_2) = T_0(z_2 - z_1) \triangleq \inf\{k \geq 1 : h_k^{(d)} \geq 0\} \tag{A.24}$$

Notice that $h_j^{(d)}|h_0^{(d)} \sim N\left(b^j(z_2 - z_1),\ 2c^2\frac{1-b^{2j}}{1-b^2}\right)$. For any n (without loss of generality, assume $n = j^2$, for some integer j),

$$\begin{aligned}
& P(T_0(z_2 - z_1) > n) \\
\leq\ & P(h_k^{(d)} < 0,\ k = 1, \cdots, n) \\
\leq\ & P(h_{jk}^{(d)} < 0,\ k = 1, \cdots, j) \\
\leq\ & \prod_{k=1}^{j} P(h_{jk}^{d} < 0) + j\ \alpha(j) \\
=\ & \prod_{k=1}^{j} \Phi\left(-b^{jk}(z_2 - z_1) \Big/ \sqrt{2c^2\frac{1-b^{2jk}}{1-b^2}}\right) + j\ \alpha(j) \\
=\ & \prod_{k=1}^{j} \Phi\left(\frac{b^{jk}(z_1 - z_2)\sqrt{1-b^2}}{\sqrt{2c^2(1-b^{2jk})}}\right) + j\ \alpha(j) \\
\leq\ & \left[\Phi\left(\frac{b^{j}(z_1 - z_2)\sqrt{1-b^2}}{\sqrt{2c^2(1-b^{2j})}}\right)\right]^j + j\ \alpha(j) \\
\leq\ & \left[\Phi\left(\frac{b^{j}(z_1 - z_2)\sqrt{1-b^2}}{c}\right)\right]^j + j\ \alpha(j).
\end{aligned} \tag{A.25}$$

Consider (A.25) in two cases:

Case 1: $z_1 - n \leq z_2 \leq x_1$. For any sufficient large n (which is equivalent to large j),

$\frac{b^j(z_1-z_2)\sqrt{1-b^2}}{c} \leq 1$. Hence $P(\tau(z_1,z_2)>n) \leq (\Phi(1))^j + j\alpha(j)$, which implies

$$\int_{z_1-n}^{z_1} P(\tau(z_1,z_2)>n)\ f(z_2;\mu_0,\sigma_0)\ dz_2 \leq (\Phi(1))^j + j\ \alpha(j)$$

Case 2: $z_2 < z_1 - n$. For sufficient large n,

$$\begin{aligned} & \int_{-\infty}^{z_1-n} P(\tau(z_1,z_2)>n)\ f(z_2;\mu_0,\sigma_0)\ dz_2 \\ \leq & \int_{-\infty}^{z_1-n} f(z_2;\mu_0,\sigma_0)\ dz_2 \\ \leq & d_3\ e^{-n} \end{aligned} \tag{A.26}$$

for some $d_3 > 0$.

A similar estimate can be obtained when $z_2 > z_1$. Putting the results together, (2.44) will hold. $\square$

Proof of **Theorem 2**:

The argument in proving Proposition 1 also applies to the non-stationary case, provided we have an error bound for the characteristic functions $\left|Ee^{i(uU_n^{(0)}+vV_n^{(0)})} - \phi(u,v)\right|$ and a Berry-Esseen bound for the cdf of $uU_n^{(0)} + vV_n^{(0)}$.

Consider $\left|Ee^{i(uU_n^{(0)}+vV_n^{(0)})} - \phi(u,v)\right|$. Based on Lemma 3 and the triangle inequality, it suffices to have

$$\begin{aligned} & \left|Ee^{i(uU_n^{(0)}+vV_n^{(0)})} - Ee^{i(uU_n^{*}+vV_n^{*})}\right| \\ = & \left|E\left\{e^{i(uU_n^{*}+vV_n^{*})}\ \left[e^{i(uw_{1n}+vw_{2n})}-1\right]\right\}\right| \\ & \text{where} \quad w_{1n} = U_n^{(0)} - U_n^*, \quad w_{2n} = V_n^{(0)} - V_n^* \\ \leq & \sqrt{E\left|e^{i(uU_n^{*}+vV_n^{*})}\right|^2}\sqrt{E\left|e^{i(uw_{1n}+vw_{2n})}-1\right|^2} \\ = & \sqrt{|E\ (e^{i(uw_{1n}+vw_{2n})}-1)(e^{-i(uw_{1n}+vw_{2n})}-1)|} \\ \leq & \sqrt{|E(e^{i(uw_{1n}+vw_{2n})}-1)| + |E(e^{-i(uw_{1n}+vw_{2n})}-1)|}. \end{aligned}$$

With the two similar terms, we only need to bound $\left|E(e^{i(uw_{1n}+vw_{2n})}-1)\right|$.

Observe that

$$\begin{aligned} U_n^{(0)} &= (Var_0 U_n^{(1)})^{-1/2}[U_n^{(1)} - EU_n^{(1)}] \\ V_n^{(0)} &= (Var_0 V_n^{(1)})^{-1/2}[V_n^{(1)} - EV_n^{(1)}] \\ U_n^* &= \sigma_{U_n}^{-1}[U_n^{(\pi)} - EU_n^{(\pi)}] \\ V_n^* &= \sigma_{V_n}^{-1}[V_n^{(\pi)} - EV_n^{(\pi)}]. \end{aligned}$$

Hence

$$w_{1n} = \frac{U_n^{(1)} - U_n^{(\pi)}}{\sqrt{Var_0 U_n}} + \frac{-E(U_n^{(1)} - U_n^{(\pi)})}{\sqrt{Var_0 U_n}} + U_n^* \left(\frac{\sigma_{U_n}}{\sqrt{Var_0 U_n}} - 1\right), \tag{A.27}$$

$$w_{2n} = \frac{V_n^{(1)} - V_n^{(\pi)}}{\sqrt{Var_0 V_n}} + \frac{-E(V_n^{(1)} - V_n^{(\pi)})}{\sqrt{Var_0 V_n}} + V_n^* \left(\frac{\sigma_{V_n}}{\sqrt{Var_0 V_n}} - 1\right). \tag{A.28}$$

Set $n_0 = b_0\ (\log n)^{\gamma_0}$ where $b_0 > 0$ and $\gamma_0 > 2$ are chosen such that the upper bound in (2.44) $c_1 \exp(-c_2 n_0^{1/2-\delta_1}) \le c_1\ n^{-1}$, hence $P(\tau(z_1) > n_0) \le c_1\ n^{-1}$. The coupling of $h^{(1)}$ and $h^{(\pi)}$ implies that on the set $\{\tau(z_1) \le n_0\}$, we have

$$U_n^{(1)} - U_n^{(\pi)} = \sum_{j=0}^{n_0-1} \left(e^{h_j^{(1)}} - e^{h_j^{(\pi)}}\right), \tag{A.29}$$

$$V_n^{(1)} - V_n^{(\pi)} = \sum_{j=0}^{n_0-1} \left(e^{h_j^{(1)}/2}\epsilon_{j+1} - e^{h_j^{(\pi)}/2}\epsilon_{j+1}\right). \tag{A.30}$$

The same argument in the proof of Lemma 3 (see the estimate for part 1 there) along with (A.27), (A.29) and Lemma 5 lead to the estimate

$$\left|E\left[(e^{i(uw_{1n}+vw_{2n})}-1)\ I_{\{\tau(z_1)\le n_0\}}\right]\right| \le O(n^{-\gamma'}) \tag{A.31}$$

where $\gamma' \in (0,1)$ can be arbitrarily close to 1.

Therefore, an error bound for the characteristic functions $\left|Ee^{i(uU_n^{(0)}+vV_n^{(0)})} - \phi(u,v)\right|$ should have the same order as the bound for $\left|Ee^{i(uU_n^*+vV_n^*)} - \phi(u,v)\right|$ obtained in Lemma 3.

Moreover, it follows from the estimates in (A.27), (A.29), Lemma 5 and the Chebyshev inequality that for $r = 8\log n$,

$$\begin{aligned} P\left(|U_n^{(0)} - U_n^*| \geq \frac{r}{8}\right) &\leq C_1 n^{-\gamma'} \\ P\left(|V_n^{(0)} - V_n^*| \geq \frac{r}{8}\right) &\leq C_1 n^{-\gamma'} \end{aligned}$$

for some $C_1 > 0$. Hence an estimate for $P\left((U_n^{(0)})^2 + (V_n^{(0)})^2 \geq \frac{r^2}{4}\right)$ similar to (A.18) can be obtained via the triangle inequalities $P\left(|U_n^{(0)}| \geq \frac{r}{4}\right) \leq P\left(|U_n^*| \geq \frac{r}{8}\right) + P\left(|U_n^{(0)} - U_n^*| \geq \frac{r}{8}\right)$ and $P\left(|V_n^{(0)}| \geq \frac{r}{4}\right) \leq P\left(|V_n^*| \geq \frac{r}{8}\right) + P\left(|V_n^{(0)} - V_n^*| \geq \frac{r}{8}\right)$.

With these estimates related to the differences w_{1n} and w_{2n}, the argument in proving Proposition 1 extends to Theorem 2. $\square$

Bibliography

Aldous, D. (1989). *Probability approximations via the Poisson clumping heuristic*. New York: Springer-Verlag.

Bhattacharya, R. and R. Rao (1976). *Normal Approximation and Asymptotic Expansions*. Wiley.

Bradley, R. (1986). Basic properties of strong mixing conditions. In E. Eberlein and M. Taqqu (Eds.), *Dependence in Probability and Statistics*, pp. 165–192.

Cheng, A., A. Gallant, C. Ji, and B. Lee (2005). A central limit theorem for computation of option prices for stochastic volatility models. Submitted to *Journal of Econometrics*.

Chernov, M. and E. Ghysels (2000). A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of option valuation. *Journal of Financial Economics 56*, 407–458.

Eisen, M., P. Spellman, P. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA 95*, 14863–14868.

Eraker, B. (2004). Do stock prices and volatility jump? reconciling evidence from spot and option prices. *Journal of Finance 54:3*, 1367–1404.

Ge, X. (2000). *Bayesian Calibration of Stochastic Volatility Models*. Ph. D. thesis, Dept. of Statistics.

Ghysels, E., A. Harvey, and E. Renault (1996). Stochastic volatility. In C. Rao and G. Maddala (Eds.), *Statistical Methods in Finance*, pp. 119–191. North-Holland.

Harrison, M. and D. Kreps (1979). Martingales and arbitrage in multiperiod security markets. *Journal of Economic Theory 20*, 381–408.

Harrison, M. and S. Pliska (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Process and Their Application 11*, 215–260.

Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies 6*, 327–343.

Ibragimov, I. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications 7*, 349–382.

Jacquier, E., N. Polson, and P. Rossi (2003). Bayesian analysis of fat-tailed stochastic volatility models with correlated errors. To appear in *J. Econometrics*.

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika 2*, 241–254.

Jones, C. (2003). The dynamics of stochastic volatility: Evidence from underlying and option markets. To appear in *Journal of Econometrics*.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An introduction to Cluster Analysis*. New York: John Wiley & Sons.

Kloeden, P. E. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Springer.

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE 78*, 1464–1479.

Lee, B., A. Cheng, and C. Ji (2004). Central limit theorems in mcmc computation of option prices with stochastic volatility models. Preprint.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. Lecam and J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 1–281. Berkeley: University of Califorinia press.

Øksendal, B. (1995). *Stochastic Differential Equations* (4th edition ed.). Springer.

Pan, J. (2002). The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics 63*.

Polson, N. and J. Stroud (2003). Bayesian inference for derivative prices. Preprint.

Reznik, M. (1968). The law of the iterated logarithm for some classes of stationary processes. *Theory of Probability and Its Applications 13*, 606–621.

Richardson, S. and P. Green (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B 59*, 731–792.

Romano, M. and N. Touzi (1997). Contingent claims and market completeness in a stochastic volatility model. *Mathematical Finance 7*, 399–412.

Stein, L. (2004). End of the beginning. *Nature 431*, 915–916.