Copy number variation and non-parametric Hidden Markov Models

Omiros Papaspiliopoulos
Universitat Pompeu Fabra
http://www.econ.upf.edu/∼omiros
Joint work with **Gareth Roberts (Warwick), Chris Holmes and Chris Yau (Oxford)**

MCMC Workshop 2009

The talk is based on three articles

Papaspiliopoulos & Roberts (2008), Retrospective MCMC for Dirichlet process hierarchical models, *Biometrika*

Papaspiliopoulos (2008), A note on posterior sampling from Dirichlet mixture models (*unpublished*)

Yau, Papaspiliopoulos, Roberts and Holmes (2008) Bayesian Nonparametric Hidden Markov Models with application to the analysis of copy-number-variation in mammalian genomes (*submitted*)

Motivating Application: Copy Number Variation (CNV)

Copy number variants are regions of the genome that can occur at variable copy number in the population. In diploid organisms, such as humans, somatic cells normally contain two copies of each gene, one inherited from each parent. However, abnormalities during the process of DNA replication and synthesis can lead to the loss or gain of DNA fragments. For example, the loss or gain of a number of tumor suppressor genes and oncogenes are known to promote the initiation and growth of cancers.

Recent studies have highlighted the complementary role of CNVs in genetic variation to SNPs. Great interest in furthering our understanding of the evolution of copy number variation and the role it may play in genetic diseases.

ROMA experiments

Microarray technology has enabled CNV across the genome to be routinely profiled using array comparative genomic hybridisation (aCGH) methods. These technologies allow DNA copy number to be measure at millions of genomic locations simultaneously allowing copy number variants to be mapped with high resolution.

Roughly speaking: immobilized genes are placed on the microarray, and strands from the same chromosome of two different subjects (case/vontrol) are extracted and colour-tagged differently. Then they are co-hybridized with the genes in the microarray and over-expression in a certain location corresponds to high copy number in the case relative to control and underexpression to low copy number.

Data: y_t, $= 1, \ldots, T$, log-ratio of hybridization levels at genomic location t. In applications $T = \mathcal{O}(10^4 - 10^6)$.
An example dataset (25)

Statistical challenge

CNV discovery amounts to detecting segmental changes in the mean levels of the DNA hybridisation intensity along the genome. However, these measurements are extremely sensitive to variations in DNA quality, DNA quantity and instrumental noise and this has lead to the development of a number of statistical methods for data analysis.

Previous approaches

One popular approach is Hidden Markov Models (HMMs) where the hidden states correspond to the unobserved copy number states at each probe location. Typically the distributions of the observations are assumed to be Gaussian or a mixture of two Gaussians or a Gaussian and uniform distribution, where the second mixture component acts to capture outliers.

However, it has been shown and our work emphasizes it that due to imperfect experimental conditions methods can be extremely sensitive to outliers, skewness or heavy tails in the actual noise process that might lead to large numbers of false copy number variants being detected.

Our contribution

- build a robust semi-parametric modelling framework.
- introduce a computational paradigm which can deal with T and be routinely used.

Effectively we provide a generic modelling/computational framework for semi-parametric product-partition modelling. Our aim is to use a so-called mixture of Dirichlet processes (MDP) for the residual distribution. Use specific representations to enhance dramatically the computations.

HMM-MDP model formulation

Let $f(y|m,z)$ be a density with parameters m and z; S_t be a Markov chain with discrete state-space $\mathcal{S} = \{1,\ldots,n\}$, transition matrix $\Pi = [\pi_{i,j}]_{i,j\in\mathcal{S}}$ and initial distribution π_0, $m = (m_1,\ldots,m_n)$ mean levels:

$$y_t \mid s_t, k_t, \mathbf{m}, \mathbf{z} \sim f\left(y_t | m_{s_t}, z_{k_t}\right),\ t = 1,\ldots,T$$

$$P(s_t = i \mid s_{t-1} = j) = \pi_{i,j}\,, i,j \in \mathcal{S}$$

$$p(k_t, u_t \mid \mathbf{w}) = \sum_{j:w_j > u_t} \delta_j(\cdot) = \sum_{j=1}^{\infty} 1[u_t < w_j]\delta_j(\cdot) \qquad (1)$$

$$z_j \mid \theta \sim H_\theta\,, j \geq 1$$

$$w_1 = v_1,\ \ w_j = v_j \prod_{i=1}^{j-1}(1 - v_i), j \geq 2$$

$$v_j \sim Be(1,\alpha)\,, j \geq 1\,,$$

Main observations

• Characterising features: structural changes in time and flexible sampling distribution at each regime. The structural changes are induced by the hidden Markov model (HMM) prior on m, as specified in the second lines in the hierarchy. The conditional distribution of y given the HMM state is specified as a mixture model in which $f(y \mid m, z)$ is mixed with respect to a random discrete probability measure $P(\mathrm{d}z)$. The last four lines in the hierarchy identify P with the Dirichlet process prior (DPP) with base measure H_θ and the concentration parameter α. Such mixture models are known as mixtures of Dirichlet process (MDP). Basically, it is an infinite mixture model with certain (convenient) structure on the mixture weights.

• We deal with a model with two levels of clustering for Y, a temporally persisting (local) clustering induced by the HMM (S) and a global clustering induced by the DPP (K).

• We have chosen a particular representation (following Walker (2007)) for DPP in terms of the allocation variables **k**, the stick-breaking weights **v**, the mixture parameters **z** and the auxiliary variables **u**. Note that **w** is a transformation of **v**. The representation of the DPP when **u** is marginalised out is well known:

$$p(k_t \mid \mathbf{w}) = \sum_{j=1}^{\infty} w_j \delta_j(\cdot)\,. \tag{2}$$

(1) clearly implies (2). (1) follows from a standard representation of an arbitrary random variable k with density p as a marginal of a pair (k, u) uniformly distributed under the curve p. When p is unimodal the representation coincides with Khinchine's theorem. The reason why we prefer the augmented representation in terms of **u** is linked to enhance dynamic programming techniques.

• A specific instance when $Y_t \in R$, f is the Gaussian density with mean $m + \mu$ and variance σ^2, $z = (\mu, \sigma^2) \in R \times R_+$, and H_Θ is a $N(0, \gamma) \times IG(a, b)$ product measure with $\Theta = (\gamma, a, b)$. Then, $E(Y_t \mid S, m) = m_t$ is a slowly varying random function driven by the HMM and the distribution of the residuals $Y_t - m_t$ is a Gaussian MDP.

Computational protocol

- The model targeted to uncover structural changes in long time series (T can be of $\mathcal{O}(10^5)$). Hence, the first requirement is that the algorithmic time scales well with T.
- Second, the algorithm should not get trapped around minor modes which correspond to confounding of local with global clustering. Informally, we would like to make moves in the high probability region of HMM configurations and then use the residuals to fit the MDP component.
- Third, we would like the algorithm to require little human intervention (i.e. Gibbs sampling vs reversible jump)

In our application we can treat Π, m, n known. We can also fix Θ. Aim to sample (K, S, U, V, Z, α). S is really the parameter of interest, the other are effectively nuisance

Block Gibbs sampling for HMM-MDP

Gibbs sampling according to the following conditional distributions:

1. $[\mathbf{s} \mid \mathbf{y}, \mathbf{u}, \mathbf{v}, \mathbf{z}]$
2. $[\mathbf{k} \mid \mathbf{y}, \mathbf{s}, \mathbf{u}, \mathbf{v}, \mathbf{z}]$
3. $[\mathbf{v}, \mathbf{u} \mid \mathbf{k}, \alpha]$
4. $[\mathbf{z} \mid \mathbf{y}, \mathbf{k}, \mathbf{s}, \mathbf{m}]$
5. $[\alpha \mid \mathbf{k}]$.

1 and 2 correspond to a joint update of **s** and **k**, by first drawing **s** from$[\mathbf{s} \mid \mathbf{y}, \mathbf{u}, \mathbf{v}, \mathbf{z}]$ and subsequently **k** from $[\mathbf{k} \mid \mathbf{y}, \mathbf{s}, \mathbf{u}, \mathbf{v}, \mathbf{z}]$. Hence, we integrate out the global allocation variables **k** in the update of the local allocation variables **s**. As a result the algorithm does not get trapped in secondary modes which correspond to mis-classification of consecutive data to Dirichlet mixture components.

1 can be seen as an update of the HMM component ("HMM update"), whereas 2-5 constitute an update of the MDP component ("MDP update").

The "MDP update" is done using a generic methodology for MDP posterior simulation (Exact Block Gibbs Sampling) which we have developed and can be applied in any other context. Although we update an infinite-dimensional variable no approximations are involved

The "HMM update" is efficiently done exploiting the structure of the DPP and a remarkable property, only shared by conditional methods
In this talk, I will not talk about the "MDP update", which is effectively a talk on its own. It is a novel algorithm based on retrospective sampling and strategic blocking of the variables.

Main result: conditional exchangeability

We can simulate exactly from $[\mathbf{s} \mid \mathbf{y}, \mathbf{u}, \mathbf{v}, \mathbf{z}]$ using a standard forward filtering/backward sampling algorithm (see for example Cappe et. al (2005)). This is facilitated by the following key result.

Proposition 1. *The conditional distribution* $[\mathbf{s} \mid \mathbf{y}, \mathbf{u}, \mathbf{v}, \mathbf{z}]$ *is the posterior distribution of a hidden Markov chain* $s_t, 1 \leq t \leq T$, *with state space* $\mathcal{S}$, *transition matrix* Π, *initial distribution* π_0, *and conditional independent observations* y_t *with conditional density,*

$$p_t(y_t \mid s_t, u_t, \mathbf{w}) = \sum_{j:w_j > u_t} f(y_t \mid m_{s_t}, z_j).$$

Proof

$$p(\mathbf{y} \mid \mathbf{s}, \mathbf{z}, \mathbf{v}, \mathbf{u}) = \sum_{\mathbf{k}} p(\mathbf{y} \mid \mathbf{s}, \mathbf{k}, \mathbf{z}) p(\mathbf{k} \mid \mathbf{w}, \mathbf{u})$$

$$= \sum_{\mathbf{k}} \prod_{t=1}^{T} f(y_t \mid m_{s_t}, z_{k_t}) p(k_t \mid u_t, \mathbf{w})$$

$$= \prod_{t=1}^{T} \sum_{j=1}^{\infty} 1[u_t < w_j] f(y_t \mid m_{s_t}, z_j) = \prod_{t=1}^{T} \sum_{j: u_t < w_j} f(y_t \mid m_{s_t}, z_j)$$

The first equality follows by standard marginalisation, where we have used the conditional independence to simplify each of the densities. The second equality follows from the conditional independence of the y_t's and the k_t's given the conditioning variables. We exploit the product structure to exchange the order of the summation and the product to obtain the third equality. The last equality is a re-expression of the previous one.

The number of terms involved in likelihood evaluations is finite a.s., since there will be a finite number of mixture components with weights $w_j > u^{*(T)} := \inf_{1 \leq t \leq T} u_t \colon\ j > j^{*(T)}$, where $j^{*(T)}$ can be identified with only partial information about the random measure $(\mathbf{z}, \mathbf{v})$ (Retrospective sampling)

However, $j^{*(T)}$ will typically grow with T. Under the prior distribution, $u^{*(T)} \downarrow 0$ almost surely as $T \to \infty$. Standard properties of the DPP imply that $j^{*(T)} = \mathcal{O}(\log T)$. This relates to the fact that the number of new components generated by the Dirichlet process grows logarithmically with the size of the data. On the other hand, it is well known that the computational cost of the forward filtering/backward sampling, when the computational cost of evaluating the likelihood is fixed, is $\mathcal{O}(T)$ (and quadratic in the size of the state space). Hence, we expect an overall computational cost $\mathcal{O}(T \log T)$ for the *exact* simulation of the hidden Markov chain in this non-parametric setup.

Numerical and methodological comparisons

In the article we argue why other parametrisations of the DPP are not appealing in this context. They either make the marginalization in the global variable update impossible, or they lead to $\mathcal{O}(T^2)$ costs.

We also provide extensive computational comparisons among different methods. Here I will present a subset of those results

Algorithmic comparisons on simulated datasets

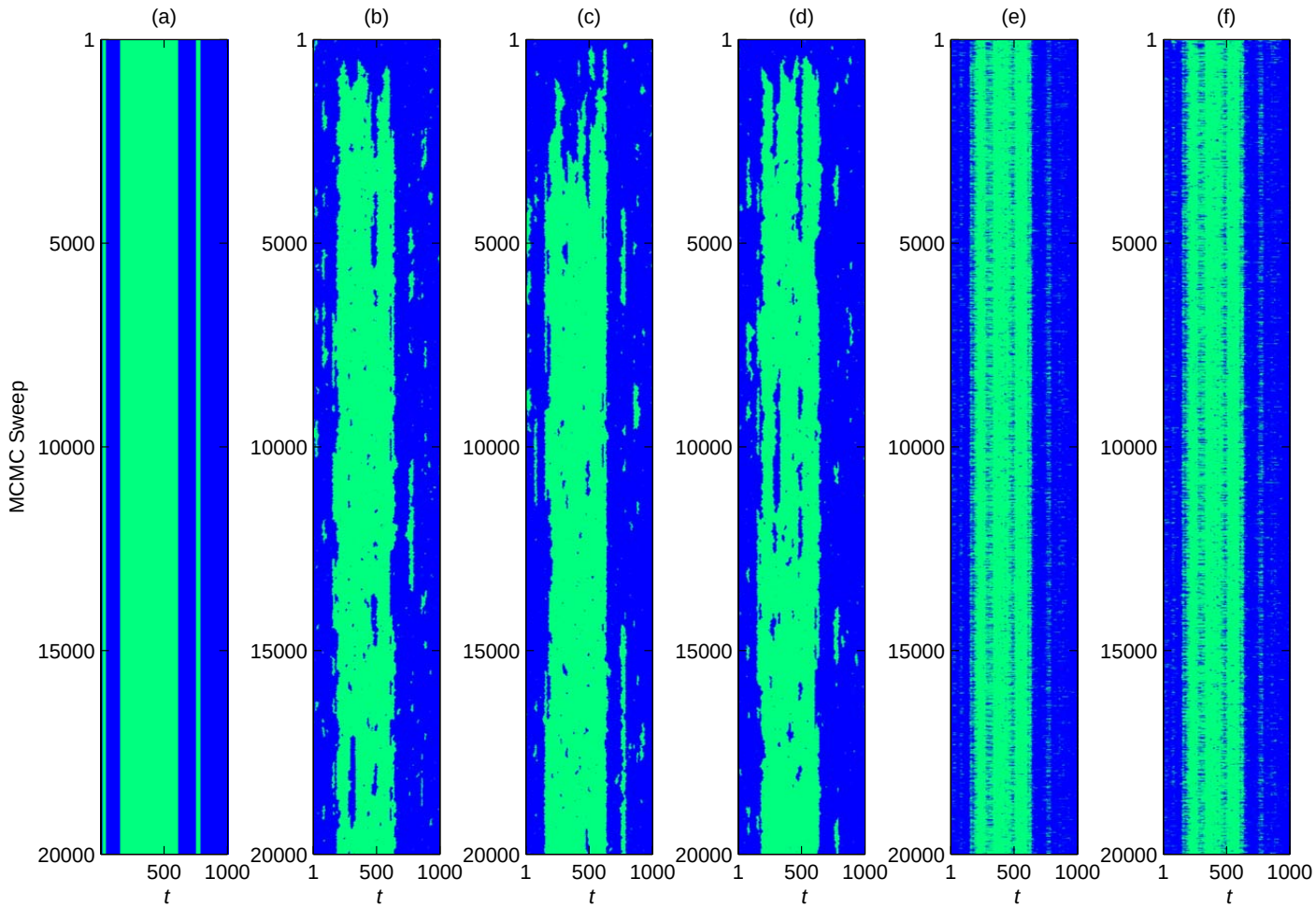


Figure: MCMC Samples of S. (a) Ground Truth, (b) Marginal Gibbs Sampler, (c) Slice Sampler with local updates, (d) Block Gibbs Sampler with local updates, (e) Slice Sampler with forward-backward updates and (f) Block Gibbs Sampler with forward-backward updates. There is a significant amount of correlation in the samples of S from the samplers employing local Gibbs updates compared to the samplers using forward-backward sampling.

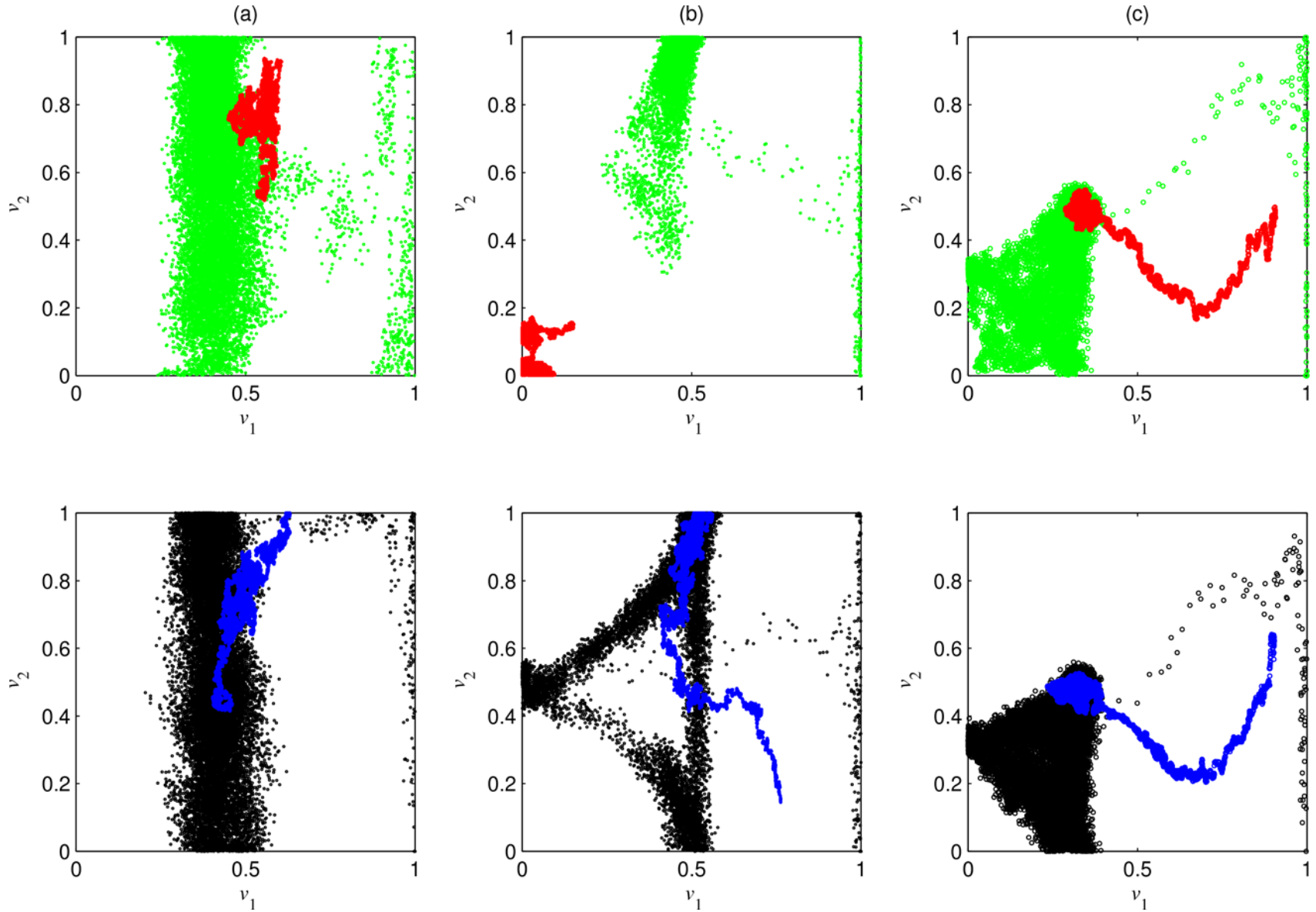


Figure: Gibbs Sampler output for (v_1, v_2). (a) *lepto 1000*, (b) *bimod 1000* and (c) *trimod 1000*. The combination of the Block Gibbs Sampler with forward-backward updating of the hidden states is able to explore the posterior distribution of v most efficiently. (Red) Slice sampler with local updates, (Green) Block Gibbs Sampler with local updates, (Blue) Slice sampler with forward-backward updates and (Black) Block Gibbs Sampler with forward-backward updates.

Model comparisons on mouse ROMA data

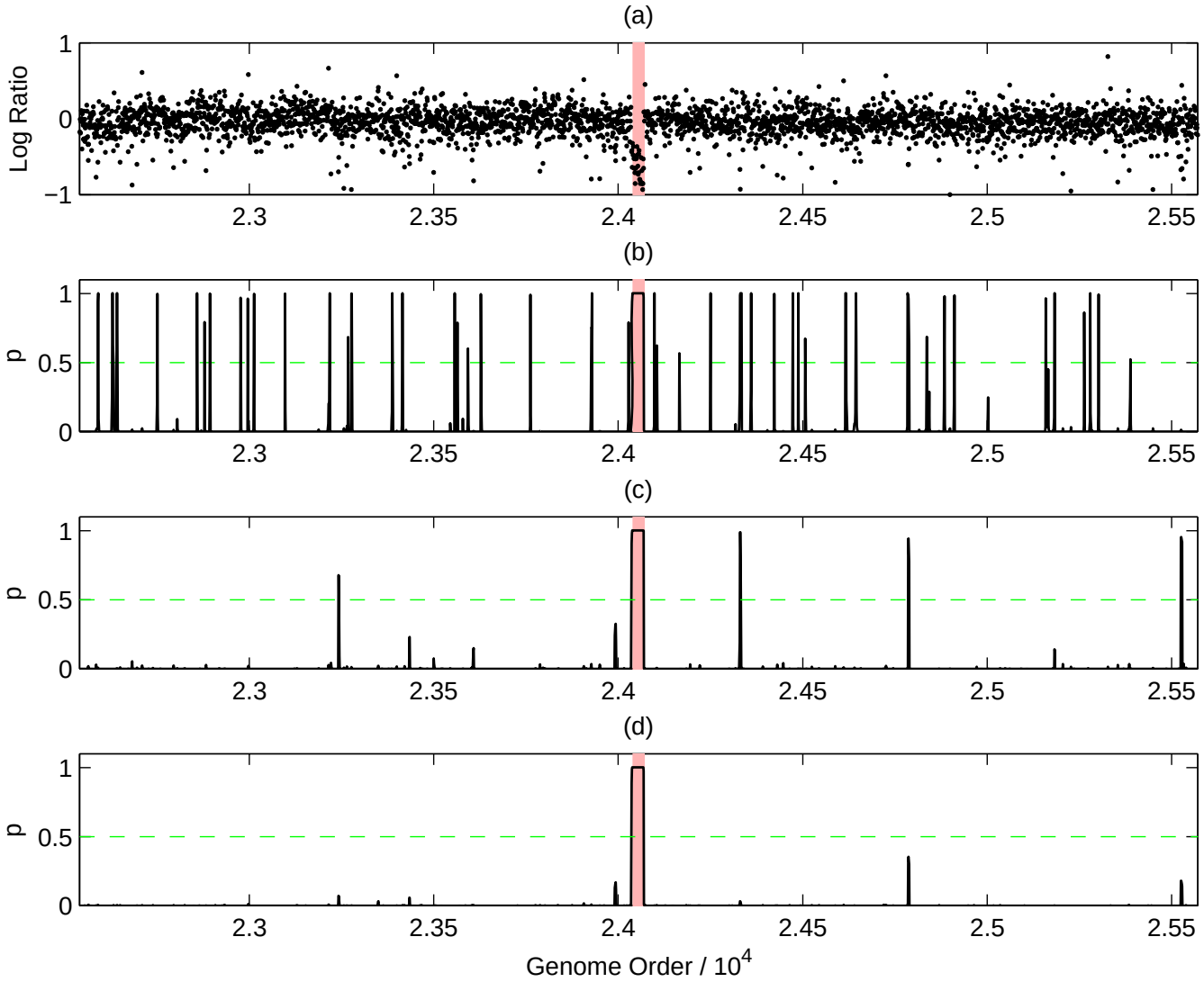


Figure: Mouse ROMA analysis. Chromosome 5. (a) The region indicated (red) contains a confirmed deletion. (b) Using the G-HMM is able to identify this known copy number variant, however, it also detects many additional copy number variants on this chromosome most of which must be false positives. (c) The R-HMM reduces the number of false positives but (d) the MDP-HMM identifies only the known copy number variant and no other copy number alterations on this chromosome.

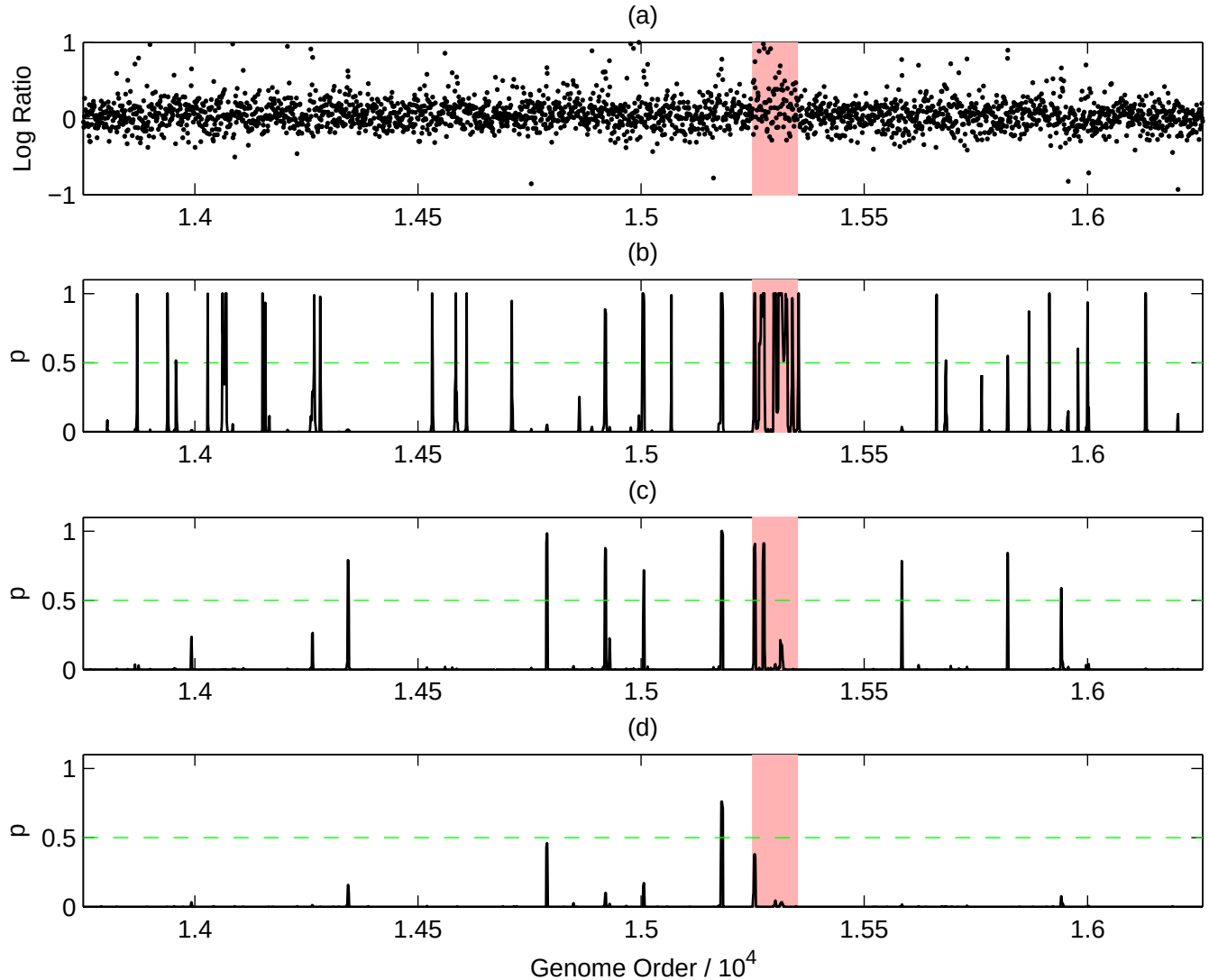


Figure: Mouse ROMA analysis. Chromosome 3. (a) The region indicated (red) contains no copy number alterations but contains SNPs that can disrupt the binding of probes on the microarray. The (b) G-HMM and (c) R-HMM produce a number of false positive copy number alteration calls in this region but (d) the MDP-HMM identifies no copy number alterations with posterior probability greater than the threshold of 0.5 in the region.

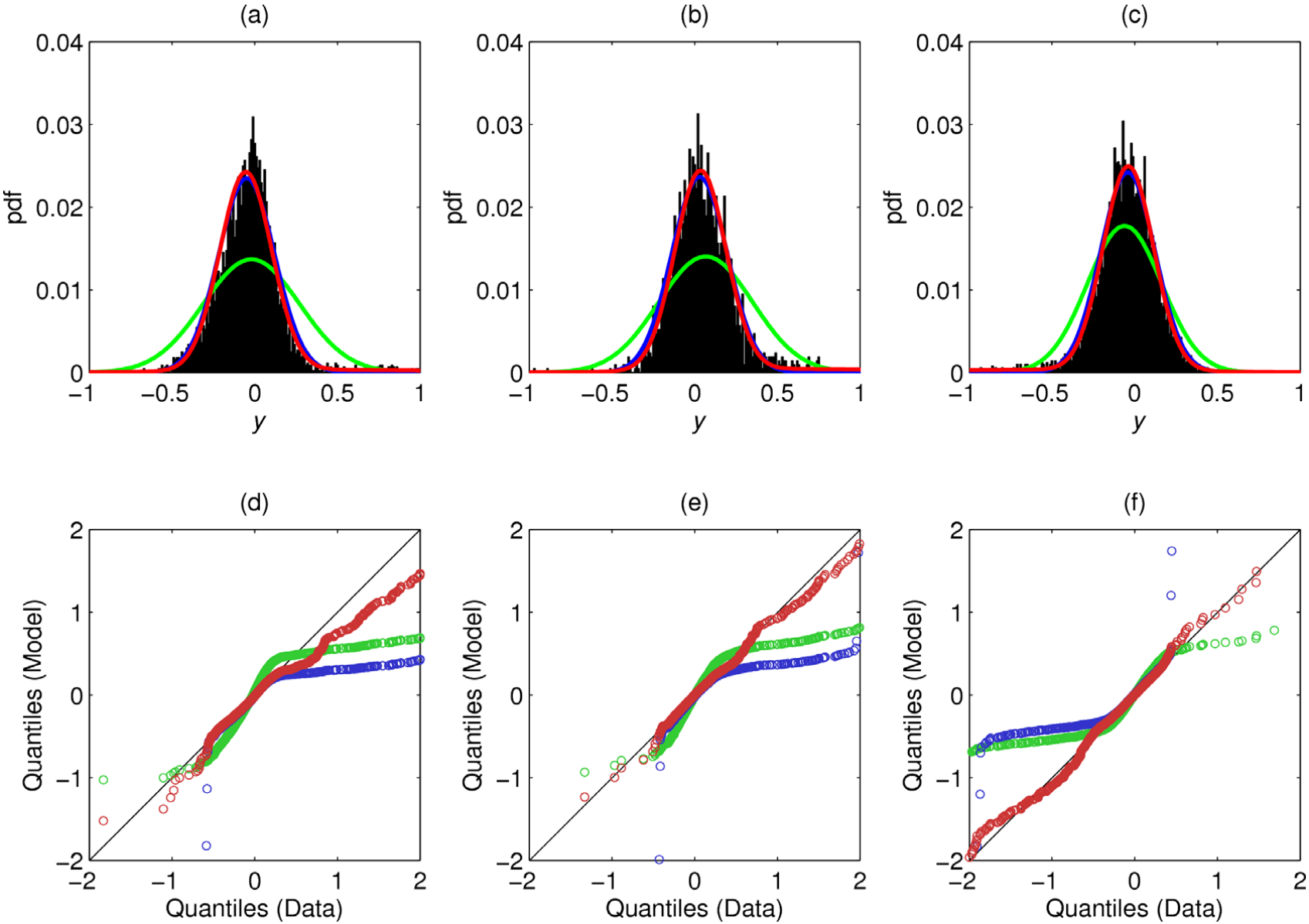
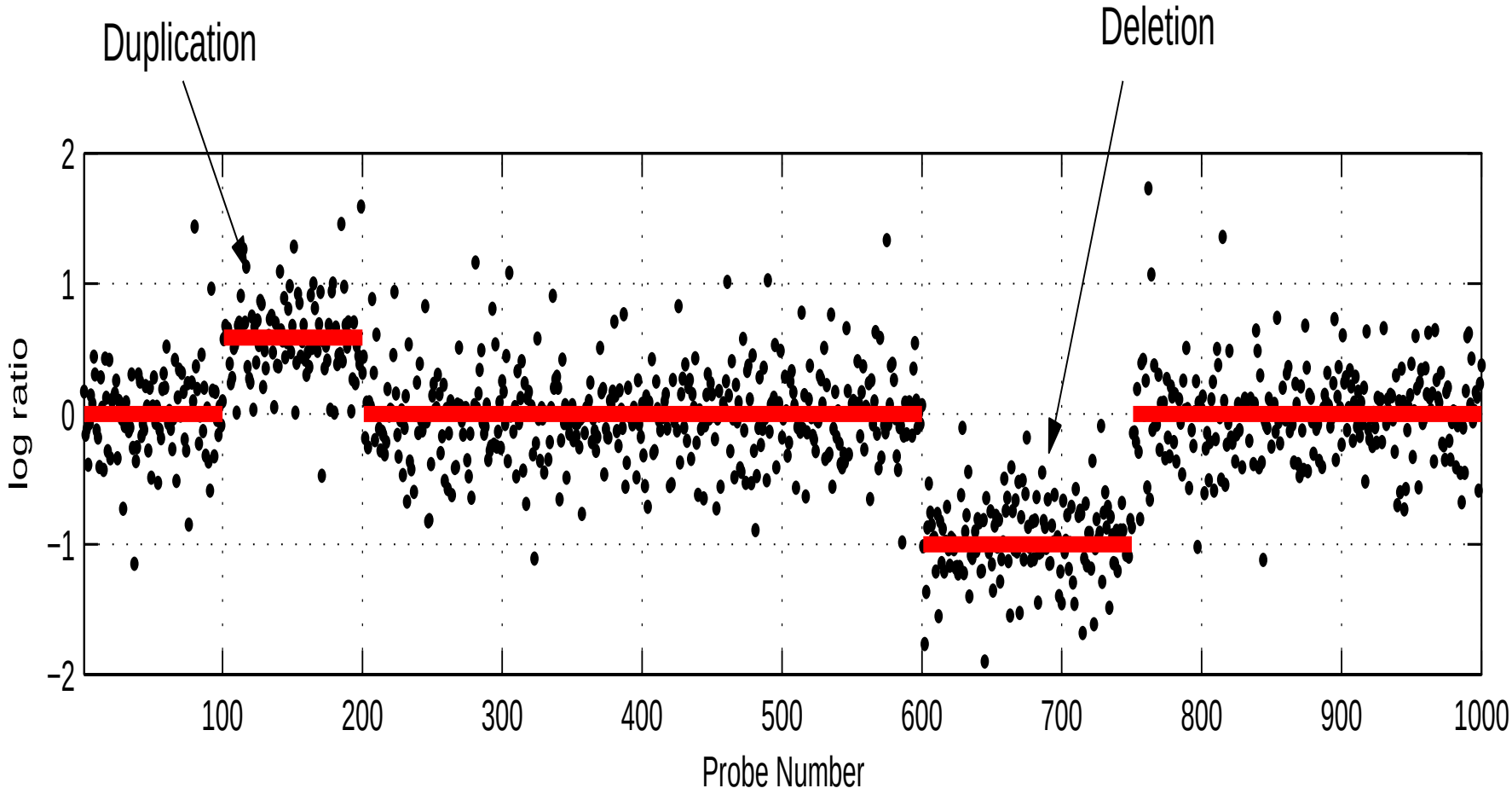


Figure: QQ-plots of predictive distributions versus ROMA data. (a, d) Chromosome 3, (b, e) Chromosome 5, (c, f) Chromosome 9. The empirical distribution of the ROMA data appears to be heavy-tailed and asymmetric. This asymmetry can lead to false detection of copy number variants by the G-HMM and R-HMM. The increased flexibility of the MDP-HMM allows this asymmetry to be capture and explains why the MDP-HMM is able to give far more accurate predictions for copy number alteration. (Red) MDP-HMM, (Green) R-HMM and (Blue) G-HMM.


Duplication
Deletion
log ratio
2
1
0
−1
−2
100
200
300
400
500
600
700
800
900
1000
Probe Number