Panagiotis Papastamoulis*, James Hensman, Peter Glaus and Magnus Rattray

Improved variational Bayes inference for transcript expression estimation

Abstract: RNA-seq studies allow for the quantification of transcript expression by aligning millions of short reads to a reference genome. However, transcripts share much of their sequence, so that many reads map to more than one place and their origin remains uncertain. This problem can be dealt using mixtures of distributions and transcript expression reduces to estimating the weights of the mixture. In this paper, variational Bayesian (VB) techniques are used in order to approximate the posterior distribution of transcript expression. VB has previously been shown to be more computationally efficient for this problem than Markov chain Monte Carlo. VB methodology can precisely estimate the posterior means, but leads to variance underestimation. For this reason, a novel approach is introduced which integrates the latent allocation variables out of the VB approximation. It is shown that this modification leads to a better marginal likelihood bound and improved estimate of the posterior variance. A set of simulation studies and application to real RNA-seq datasets highlight the improved performance of the proposed method.

Keywords: Kullback-Leibler divergence; marginal likelihood bound; generalized Dirichlet distribution; BitSeq; mixture model.

Availability: C++ source code is hosted on github.com/mqbssppe/gen_dir_vb.git.

*Corresponding author: Panagiotis Papastamoulis, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK, e-mail: panagiotis.papastamoulis@manchester.ac.uk

James Hensman: University of Sheffield, The Sheffield Institute for Translational Neuroscience, 385A Glossop Road, Sheffield, S10 2HQ, UK

Peter Glaus and Magnus Rattray: University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK

1 Introduction

The transcriptome is the set of all transcripts in a studied organism or in a specific cell, at a given developmental stage or biological condition. Determining the transcriptome is an essential task in molecular biology because it reveals functional elements of the genome. The aim of transcriptomics is either to construct the transcriptome using de novo assembly, or to quantify the expression of transcripts using a reference assembled transcriptome. RNA-seq (Marioni et al., 2008; Mortazavi et al., 2008) is an emerging next generation sequencing technology, surpassing the previously adopted microarray experiments (Wang et al., 2009).

The RNA-seq procedure results in a dataset of short reads, consisting of nucleotide sequences. These reads are aligned to the reference genome or transcriptome using bioinformatics tools such as Bowtie (Langmead et al., 2009) or TopHat (Trapnell et al., 2009). The task of inferring transcript abundances given the aligned short reads would be straightforward if these alignments were unique. However, during the process of transcription, most genes can be alternatively spliced into various transcripts which share specific parts of their sequence (exons). Different alleles of the same gene also share most of their sequence. Consequently, many reads map to several transcripts of interest and the estimation of transcript abundance has to be treated probabilistically, since the origin of each read is unknown.

In statistical terms, the problem of inferring relative transcript expression reduces to estimating the weights of a mixture model, assuming that the transcriptome is known. Li et al. (2010) applied a maximum likelihood approach, using the expectation-maximization (EM) algorithm (Dempster et al., 1977). The Bayesian approach was followed by Katz et al. (2010), Turro et al. (2011) and Glaus et al. (2012), via Markov chain Monte Carlo (MCMC) sampling. However, the high dimensionality of RNA-seq datasets imposes certain new

inferential difficulties, making convergence of MCMC methods a time consuming task. On the other hand, approximate inference using variational Bayesian methods (Jordan et al., 1999) offers an attractive alternative which has also been applied to transcript abundance inference (Hensman et al., 2012, 2013; Nariai et al., 2013). The posterior distribution is approximated by another one which is available in analytical form. At the same time, a lower bound of the marginal likelihood is provided which may be useful for model selection.

Data augmentation (Tanner and Wong, 1987) is common practice in mixture models, considered both by maximum likelihood and Bayesian approaches. Standard implementation of VB methodology (Jordan et al., 1999; Beal and Ghahramani, 2003) approximates the joint posterior distribution of mixture weights and missing data, rather than the actual posterior, using an EM-like procedure (VBEM). The approximation is based on a factorization assumption of the latent variables and model parameters. Recent advances, based on the same factorization assumption, indicate a significant inference speed-up. Nariai et al. (2013) apply VBEM to RNA-seq data showing a speed-up over standard EM. Hensman et al. (2012) have developed a new algorithm based on Riemannian gradient descent. Their method is an order of magnitude faster than traditional VB implementation, which is itself often faster than MCMC. This algorithm provides a significant speed-up in comparison to VBEM for transcript quantification (Hensman et al., 2012, 2013).

In this paper we are concerned with the quality of VB inference. In particular, we show that it is better to approximate the actual posterior distribution, rather than the joint posterior of model parameters and latent variables. The proposed methodology builds upon the BitSeq model (Glaus et al., 2012) and exploits the solution of standard VB. An optimization is performed over a class of distributions that share the same mean as the VB solution, but the variance is different. Two different parameterizations are considered: the first one forces the approximating distribution to remain inside the Dirichlet family, while the second relaxes this assumption by using the generalized Dirichlet (Wong, 1998) family. It is proved that the new method can provide a tighter lower bound of the log-marginal likelihood. Moreover, a set of simulation studies highlights that the new approach leads to a better posterior variance estimate.

The rest of the paper is organized as follows. Section 2 introduces some notation and describes the mixture model. The standard VB implementation is briefly reviewed in Section 2.1. In Section 2.2 a better bound is constructed and the optimization problem is stated in its general form. Two parameterizations for the optimization problem are given in Section 2.3, resulting in the Dirichlet and generalized Dirichlet distributions, respectively. The methodology is illustrated in a simulation study and applied to real RNA-seq datasets in Sections 3.1 and 3.2. Finally, an application under a differential expression setup is given at Section 3.3. The paper concludes in Section 4 with a discussion.

2 Methods

Let $\mathbf{x}=(x_1, \dots, x_n)$ denote n independent observations identically distributed according to a finite mixture of $K>1$ known distributions, $f_1, \dots, f_K$, that is,

$$x_i \sim \sum_{k=1}^K \theta_k f_k(x_i), \quad i=1, \dots, n. \quad (1)$$

Let $\boldsymbol{\theta}=(\theta_1, \dots, \theta_{K-1}) \in \Theta_K := \{\boldsymbol{\theta}_k > 0, k=1, \dots, K-1: \sum_{k=1}^{K-1} \theta_k < 1\}$ denote the unknown weights with $\theta_K := 1 - \sum_{k=1}^{K-1} \theta_k$. Let $\mathbf{z}_i=(z_{i1}, \dots, z_{iK})$ be the latent vector which assigns the i -th observation to one of the components, that is, $x_i | (z_{ik}=1) \sim f_k(x_i)$, with $\mathbf{z}_i | \boldsymbol{\theta} \sim \text{Multinomial}(1; \theta_1, \dots, \theta_K)$, independently for $i=1, \dots, n$, where Multinomial denotes the multinomial distribution. The latent variable structure of (1) is exploited both by the EM algorithm (Dempster et al., 1977) as well as the Gibbs sampler (Gelfand and Smith, 1990). The reader is referred to McLachlan and Peel (2000) and Marin et al. (2005) for an overview of frequentist and Bayesian approaches to finite mixture modelling.

In terms of transcript estimation, x_i denotes a short read, while $f_k(x_i)$ corresponds to the probability of read i aligning at some position of transcript k . Since we assume a known transcriptome, $\{f_k\}_{k=1}^K$ are known as well. These alignment probabilities are precomputed using the methodology described in Glaus et al.

(2012). Each read can potentially map to any of the K transcripts. Let now θ_k denote the (unknown) relative expression of transcript $k=1, \dots, K$. Then, model (1) describes the overall probability of read $i=1, \dots, n$ being aligned to any of the K transcripts.

Under a Bayesian setup, let $p(\boldsymbol{\theta})$ be the prior distribution, which in our context is a Dirichlet distribution, that is,

$$\boldsymbol{\theta} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K),$$

where $\alpha_k > 0$, $k=1, \dots, K$, denote a set of known hyperparameters. The marginal likelihood, defined as

$$m(\mathbf{x}) := \int_{\Theta_K} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta_K} \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (2)$$

is an important quantity to estimate because it allows for model selection. MCMC estimation of $m(\mathbf{x})$ is possible but not straightforward in general settings (see for example Chib, 1995; Frühwirth-Schnatter, 2006, p. 139).

2.1 Standard variational approximation

VB methods can be constructed by considering a lower-bound (L) of $\log m(\mathbf{x})$, by performing a free-form minimization of the Kullback-Leibler (KL) divergence $KL(q||p)$ between an approximating distribution q and the true posterior distribution p . We write

$$\log m(\mathbf{x}) = L + KL(q||p). \quad (3)$$

From this expression, it is clear that minimization of the KL divergence is equivalent to maximization of the lower bound, since the KL is non-negative. According to the standard VB methodology (Jordan et al., 1999; Beal and Ghahramani, 2003; Hensman et al., 2012; Nariai et al., 2013), the joint augmented posterior $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$ is approximated by another distribution $q(\boldsymbol{\theta}, \mathbf{z})$. However, if no further assumptions are made, this maximization results to $q(\boldsymbol{\theta}, \mathbf{z}) = p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$, that is, the true posterior. Obviously, that will not simplify the situation at all, since in this case the knowledge of the normalizing constant (i.e., the marginal likelihood) is required.

To make the problem tractable the optimization is done within a restricted family of distributions

$$\mathcal{G} = \{g(\boldsymbol{\theta}, \mathbf{z}) = g(\boldsymbol{\theta})g(\mathbf{z}) : g(\mathbf{z}) = \prod_{i=1}^n \prod_{k=1}^K \phi_{ik}^{z_{ik}}\}, \quad (4)$$

where $\boldsymbol{\phi} := (\phi_{ik} : i=1, \dots, n, k=1, \dots, K)$ are the variational parameters. The independence assumption between $\mathbf{z}$ and $\boldsymbol{\theta}$ in (4) is common when applying VB methodology to mixture models (Corduneanu and Bishop, 2001; Beal and Ghahramani, 2003; McGrory and Titterton, 2007; Hensman et al., 2012; Nariai et al., 2013). Furthermore, notice that only the distribution of $\mathbf{z}$ is parameterized, whereas the approximate distribution of $\boldsymbol{\theta}$ will be explicitly derived as a by-product of the analytical evaluation of the bound.

Hensman et al. (2012, 2013) derived a bound to the logarithm of (2) with respect to the distributions considered in (4) in the following manner. By Jensen's inequality

$$\begin{aligned} \log p(\mathbf{x}|\boldsymbol{\theta}) &= \log \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \log \sum_{\mathbf{z} \in \mathcal{Z}} \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})q(\mathbf{z})}{q(\mathbf{z})} = \log \mathbb{E}_{q(\mathbf{z})} \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \\ &\geq \mathbb{E}_{q(\mathbf{z})} \left\{ \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \right\} = \sum_{i=1}^n \sum_{k=1}^K \phi_{ik} \{ \log f_k(x_i) + \log \theta_k - \log \phi_{ik} \}, \end{aligned}$$

for all $q(\mathbf{z}) \in \mathcal{G}$, $\boldsymbol{\theta} \in \Theta_K$. Now substitute the last inequality to (2) to obtain that

$$m(\mathbf{x}) \geq \int_{\Theta_K} \exp \left(\sum_{i=1}^n \sum_{k=1}^K \phi_{ik} \{ \log f_k(x_i) + \log \theta_k - \log \phi_{ik} \} \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (5)$$

Figure 1 Simulated datasets of Section 3.1: Kernel density estimates of $\theta_k | \mathbf{x}$ according to: MCMC (—), standard VB (---), Dirichlet (---), generalized Dirichlet (—). $k=1,2,3$ for dataset A, B and $k=1,5,9$ for dataset C. The vertical lines denote the estimated MCMC (—) and (common) VB means (---).

The analytical evaluation of the integral (5) leads to

$$\log m(\mathbf{x}) \geq \mathcal{L}_1(\boldsymbol{\phi}) := c + \sum_{k=1}^K \log \Gamma \left(\alpha_k + \sum_{i=1}^n \phi_{ik} \right) + \sum_{i=1}^n \phi_{ik} \{ \log f_k(x_i) - \log \phi_{ik} \} \quad (6)$$

with c and $\Gamma(\cdot)$ denoting a known constant (not depending on $\boldsymbol{\phi}$) and the gamma function, respectively. Moreover, the analytical integration of (5) implies that the approximate distribution for θ is

$$q(\boldsymbol{\theta}) = \mathcal{D}(\gamma_k; k=1, \dots, K), \quad (7)$$

where $\gamma_k := \alpha_k + \sum_{i=1}^n \phi_{ik}$. Finally, $\mathcal{L}_1(\boldsymbol{\phi})$ is maximized with respect to $\boldsymbol{\phi}$ usually using EM-type updates (VBEM). The proposed algorithm of Hensman et al. (2012, 2013) maximizes (6) using the natural gradient.

Let $\hat{p}(\theta_k | \mathbf{x})$ denotes an estimate of the marginal posterior density of transcript $k=1, \dots, K$, arising from long MCMC runs which will be considered as the ground truth. Figure 1 displays a comparison of $\hat{p}(\theta_k | \mathbf{x})$ and the VB approximation in (7) (red lines), based on the simulation studies in Section 3. It is obvious that this approach exhibits good performance in terms of posterior means (black and red dashed vertical lines), but it leads to variance underestimation. The distribution in (7) is optimal in terms of minimizing the KL divergence between the *joint* posterior $p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{x})$ and the distributions considered in (4). However, this does not mean that it is the “best” Dirichlet approximation of the posterior $p(\boldsymbol{\theta} | \mathbf{x})$, as the green curves suggest. A way to find a better approximation is described at the next section.

2.2 Approximating the non-augmented posterior

Independence among the latent variables ($\mathbf{z}$) and the weights ($\boldsymbol{\theta}$) of the mixture model is a convenient assumption from the variational inference point of view but, on the other hand, it is an oversimplifying one.

Here, we are aiming to relax this assumption by integrating out the latent variables from the variational approximation. First, we show that the standard variational bound ($\mathcal{L}_1$) is worse compared to a new bound ($\mathcal{L}_2$) arising by marginalization of the same approximating distribution over the latent space. This is proven in the following Proposition.

Proposition 2.1 *Let $q(\theta)$ defined as in (7) and*

$$\mathcal{L}_2 := \int_{\Theta_K} \{\log p(\mathbf{x}|\theta) + \log p(\theta) - \log q(\theta)\} q(\theta) d\theta. \quad (8)$$

Then,

$$\log m(\mathbf{x}) \geq \mathcal{L}_2 \geq \mathcal{L}_1, \quad (9)$$

and the equality holds if and only if $q(\theta, \mathbf{z}) = p(\theta, \mathbf{z}|\mathbf{x})$, $\forall \theta, \mathbf{z}$.

Proof. It is straightforward to show that (8) is a lower bound of $\log m(\mathbf{x})$. Indeed, by Equation (3), the lower bound between q and p is expressed as

$$\begin{aligned} L &= \log m(\mathbf{x}) - \text{KL}(q(\theta) \| p(\theta|\mathbf{x})) \\ &= \log m(\mathbf{x}) - \int \{\log q(\theta) - \log p(\mathbf{x}|\theta) - \log p(\theta) + \log m(\mathbf{x})\} q(\theta) d\theta \\ &= \int_{\Theta_K} \{\log p(\mathbf{x}|\theta) + \log p(\theta) - \log q(\theta)\} q(\theta) d\theta =: \mathcal{L}_2. \end{aligned} \quad (10)$$

Next, we prove the second inequality in expression (9). By the log-sum inequality,

$$q(\theta) \log \frac{q(\theta)}{p(\theta|\mathbf{x})} = \left(\sum_{\mathbf{z}} q(\theta, \mathbf{z}) \right) \log \frac{\sum_{\mathbf{z}} q(\theta, \mathbf{z})}{\sum_{\mathbf{z}} p(\theta, \mathbf{z}|\mathbf{x})} \leq \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log \frac{q(\theta, \mathbf{z})}{p(\theta, \mathbf{z}|\mathbf{x})},$$

$\forall \theta \in \Theta$. Integrating both sides of the last inequality we obtain that

$$\begin{aligned} \int q(\theta) \log \frac{q(\theta)}{p(\theta|\mathbf{x})} d\theta &\leq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log \frac{q(\theta, \mathbf{z})}{p(\theta, \mathbf{z}|\mathbf{x})} d\theta \Leftrightarrow \\ \text{KL}(q(\theta) \| p(\theta|\mathbf{x})) &\leq \text{KL}(q(\theta, \mathbf{z}) \| p(\theta, \mathbf{z}|\mathbf{x})). \end{aligned} \quad (11)$$

Now, substitute (11) into (10) and conclude that $\mathcal{L}_2 \geq \mathcal{L}_1$. Recall that the log-sum inequality holds as equality if and only if $q(\theta, \mathbf{z}) = p(\theta, \mathbf{z}|\mathbf{x})$, $\forall \theta, \mathbf{z}$ and this completes the proof. $\square$

It should be clear that $\mathcal{L}_2$ in Proposition 2.1 is just a better estimate of the lower bound, based on the same approximation of $\theta|\mathbf{x}$. The difference with the initial bound (6) is that we marginalize over $\mathbf{z}$ and not take into account the factorization assumption (4). Nevertheless, the main challenge is to improve this lower bound by finding a better approximating distribution inside a defined class of distributions, instead of just providing a tighter lower bound. This is discussed in the sequel. Let $\mathcal{F}$ denotes any subset/family of distributions with $q(\theta) \in \mathcal{F}$. Expression (11) implies that

$$\min_{f \in \mathcal{F}} \text{KL}(f(\theta) \| p(\theta|\mathbf{x})) \leq \text{KL}(q(\theta, \mathbf{z}) \| p(\theta, \mathbf{z}|\mathbf{x})). \quad (12)$$

Hence, a better bound of the marginal likelihood can be constructed in the following way: At first, a set $\mathcal{F}$ which contains $q(\theta)$ should be specified. Then, an optimization will find the best approximating distribution.

Let $\Delta_{\mathcal{F}}$ denotes the parameter space of $\mathcal{F}$ and assume that $\delta \in \Delta_{\mathcal{F}}$. Considering now that the approximation f is varying in the set $\mathcal{F}$, the bound in Equation (8) is extended to

$$\mathcal{L}_2(\delta) = \int_{\Theta_K} \{\log p(\mathbf{x}|\theta) + \log p(\theta) - \log f(\theta; \delta)\} f(\theta; \delta) d\theta, \quad (13)$$

which is a bound to $\log m(x)$, for all $\delta \in \Delta_{\mathcal{F}}$. We stress that (13) cannot be computed directly even for fixed δ . However, (13) can be approximated via stochastic approximation, since it is expressed as the mean value of the random variable $g(\theta) := \log p(\mathbf{x}|\theta) + \log p(\theta) - \log f(\theta; \delta), \theta \sim f(\cdot; \delta)$. So, our objective function is written as:

$$\max_{\delta \in \Delta_{\mathcal{F}}} \mathcal{J}_2(\delta) = \max_{\delta \in \Delta_{\mathcal{F}}} \mathbb{E}_{\delta} g(\theta), \quad \theta \sim f(\cdot; \delta) \in \mathcal{F}. \quad (14)$$

2.3 Constructing $\mathcal{F}$

Having in mind that the best variational approximation targeting the joint posterior is the Dirichlet distribution in (7), an obvious choice for $\mathcal{F}$ would be (a subset of) the Dirichlet family of distributions. However, it will prove useful to take into account an even broader family as well, that is, the generalized Dirichlet family of distributions (Wong, 1998, 2010). The VB solution (7) can be expressed as a generalized Dirichlet distribution:

$$q(\theta) = \mathcal{D}(\gamma_1, \dots, \gamma_K) \equiv \mathcal{GD}(\gamma_1, \dots, \gamma_{K-1}; \gamma_1^+, \dots, \gamma_{K-1}^+), \quad (15)$$

where $\gamma_\ell^+ := \sum_{j=\ell+1}^K \gamma_j$, $\ell=1, \dots, K-1$.

Next we define two different parameterizations of $\mathcal{F}$ in Equation (14), denoted by $\mathcal{F}_{\mathcal{G}}$ and $\mathcal{F}_{\mathcal{GG}}$ with $\mathcal{F}_{\mathcal{G}} \subset \mathcal{F}_{\mathcal{GG}}$. This is done by considering transformations of the parameters $(\gamma, \gamma^+) \rightarrow (\tilde{\gamma}, \tilde{\gamma}^+)$. To simplify the optimization we keep the same mean as the original VB distribution (7). This is reasonable due to the fact that, according to our simulation study (as well as many others not reported here), this distribution is quite accurate in estimating the posterior means.

The first transformation is based on only one variable $\delta \in \mathbb{R}$, that is,

$$(\tilde{\gamma}_k, \tilde{\gamma}_k^+) = (e^\delta \gamma_k, e^\delta \gamma_k^+), \quad k=1, \dots, K-1.$$

Note here that this transformation implies that

$$\begin{aligned} \tilde{\gamma}_{k+1} + \tilde{\gamma}_{k+1}^+ &= e^\delta \gamma_{k+1} + e^\delta \gamma_{k+1}^+ = e^\delta (\gamma_{k+1} + \gamma_{k+1}^+) = e^\delta \gamma_k^+ \\ &= \tilde{\gamma}_k^+, \quad \forall k=1, \dots, K-1 \end{aligned}$$

hence, for all δ the resulting distribution is still a member of the subset of the Dirichlet family:

$$\mathcal{F}_{\mathcal{G}} := \{ \mathcal{D}(e^\delta \gamma_1, \dots, e^\delta \gamma_K) : \delta \in \mathbb{R} \}. \quad (16)$$

The second transformation relaxes the restriction of remaining inside the Dirichlet family. Now $\delta = (\delta_1, \dots, \delta_{K-1})$,

$$(\tilde{\gamma}_k, \tilde{\gamma}_k^+) = (e^{\delta_k} \gamma_k, e^{\delta_k} \gamma_k^+), \quad k=1, \dots, K-1 \quad (17)$$

resulting to the following subset of the generalized Dirichlet family:

$$\mathcal{F}_{\mathcal{GG}} := \{ \mathcal{GD}(e^{\delta_1} \gamma_1, \dots, e^{\delta_{K-1}} \gamma_{K-1}; e^{\delta_1} \gamma_1^+, \dots, e^{\delta_{K-1}} \gamma_{K-1}^+) : \delta_k \in \mathbb{R}, k=1, \dots, K \}. \quad (18)$$

Note that the number of parameters in (16) and (18) equals one and $K-1$, respectively. Moreover, for all $f \in \mathcal{F}_{\mathcal{GG}}$ it holds that,

$$\mathbb{E}_{\theta} = \frac{\tilde{\gamma}_k}{\tilde{\gamma}_k + \tilde{\gamma}_k^+} \prod_{j=1}^{k-1} \frac{\tilde{\gamma}_j}{\tilde{\gamma}_j + \tilde{\gamma}_j^+} = \frac{e^{\delta_k} \gamma_k}{e^{\delta_k} \gamma_k + e^{\delta_k} \gamma_k^+} \prod_{j=1}^{k-1} \frac{e^{\delta_j} \gamma_j}{e^{\delta_j} \gamma_j + e^{\delta_j} \gamma_j^+} = \frac{\gamma_k}{\gamma_1 + \gamma_1^+} = \frac{\gamma_k}{\sum_{j=1}^K \gamma_j},$$

$\forall k=1, \dots, K$, while the same remains true for $f \in \mathcal{F}_g$ as well, since $\mathcal{F}_g \subset \mathcal{F}_{gg}$. Consequently, both families (16) and (18) contain distributions having the same means as the distribution $q(\theta)$ in (7). Finally, notice that when $K=2$ both parameterizations are the same, as in such a case a generalized Dirichlet distribution degenerates to a Dirichlet. In order to maximize (14) under parameterizations (16) or (18) the following stochastic approximation algorithm was implemented.

2.4 Stochastic approximation algorithm

Notice that the objective function in Equation (14) can be estimated for any given value of δ . Suppose that θ_m , $m=1, \dots, M$ is a random sample from $f(\cdot, \delta)$. Then,

$$\hat{L}_M(\theta, \delta) := \frac{1}{M} \sum_{m=1}^M (\log p(\mathbf{x} | \theta_m) + \log p(\theta_m) - \log f(\theta_m; \delta)), \quad (19)$$

is an unbiased estimator of $L(\delta)$. The problem of optimizing a function which is only observed under the presence of noise can be handled using Stochastic Approximation algorithms (Spall, 1992, 2000). Let $d := \text{card}(\delta)$ denote the number of parameters, that is either equal to 1 or $K-1$.

1. Set $t=0$ and give some initial values $\delta^{(t)}$.
2. For $t=1, 2, \dots$
 - a. Compute a gradient approximation $\hat{\lambda} = \hat{\lambda}(\theta, \delta^{(t-1)})$ of ∇L :
 - i. Simulate $\mathbf{b} = (b_1, \dots, b_d)$, $b_j \sim \mathcal{U}_{\{-1,1\}}$, independently for $j=1, \dots, d$.
 - ii. Simulate $\theta_m \sim f(\cdot; \delta^{(t-1)})$, $m=1, \dots, M$ and compute

$$L^+ := \hat{L}_M(\theta, \delta^{(t-1)} + c_{t-1} \mathbf{b}).$$
 - iii. Simulate $\theta_m \sim f(\cdot; \delta^{(t-1)})$, $m=1, \dots, M$ and compute

$$L^- := \hat{L}_M(\theta, \delta^{(t-1)} - c_{t-1} \mathbf{b}).$$
 - iv. Set

$$\hat{\lambda}(\theta, \delta^{(t-1)}) = \frac{L^+ - L^-}{2c_{t-1} \mathbf{b}}$$
 - b. Update the parameters: $\delta^{(t)} = \delta^{(t-1)} + a_{t-1} \hat{\lambda}$.

The “gain” sequences $\{a_t\}$, $\{c_t\}$, $t=1, 2, \dots$, satisfy the conditions described in Spall (1992, 2000), which guarantee strong convergence of $\delta^{(t)}$ to the optimal value, as $t \rightarrow \infty$. Following Spall (2000) we set:

$$a_t = \frac{a}{(t+A)^a}, \quad \text{and} \quad c_t = \frac{c}{t^\gamma}, \quad t=1, 2, \dots, \quad (20)$$

for some $a, A, c, \gamma > 0$. Steps 2.(a).i – 2.(a).iv implement the “simultaneous perturbation” method of Spall (1992, 2000), in order to compute a Monte Carlo based approximation of the gradient ∇L . This technique allows for simultaneous changes at all elements of the parameter vector δ , and considerably reduces the number of objective function evaluations per iteration, compared to a standard finite difference approach (2 vs $2d$). $\mathcal{U}$ refers to the discrete uniform distribution on the set $\{-1, 1\}$ (or, in other words, a zero-mean Bernoulli ± 1 random variable). Under a slight abuse of notation, the fraction at step 2.(a).iv stands for the division of the nominator by each element of the vector in the denominator.

According to Hutchison and Spall (2004), the establishment of simple stopping criteria is not an easy task for the finite sample practice of stochastic approximation algorithms. For our application, a good performance was obtained using the following heuristic. The moving average of the parameter values is computed every s iterations, that is:

$$\bar{\boldsymbol{\delta}}^{(\tau)} = \frac{1}{S} \sum_{t=\tau-S+1}^{\tau} \boldsymbol{\delta}^{(t)}, \quad \text{where } \tau = s, 2s, 3s, 4s, \dots$$

Then, given a pre-specified positive integer v , denote by

$$\hat{L}^{(\tau)} := \hat{L}_q(\boldsymbol{\theta}, \bar{\boldsymbol{\delta}}^{(\tau)}), \quad \text{for } \tau = vs, (v+1)s, (v+2)s, \dots,$$

the estimate of the objective function (19) evaluated at $\bar{\boldsymbol{\delta}}^{(\tau)}$. Let now

$$V^{(\tau)} := \{\hat{L}^{(\tau-v+j)}, j=1, \dots, v\}, \quad \text{for } \tau = vs, (v+1)s, (v+2)s, \dots,$$

denote the set of the last v evaluations. Finally, let

$$R^{(\tau)} := \{\text{sign}(V_{j+1}^{(\tau)} - V_j^{(\tau)}), j=2, \dots, v\}.$$

If at iteration $\tau \in \{vs, (v+1)s, (v+2)s, \dots\}$, the number of runs at sign vector $R^{(\tau)}$ equals to $v-1$, the algorithm is terminated.

2.5 Practical implementation

Before proceeding to the applications, we describe the parameter settings we used in order to obtain a satisfactory performance of the stochastic approximation algorithm. The learning rates in Equations (20) play a key role in the practical implementation of the algorithm. However, we should underline the fact that there are no general rules to adjust their parameters. According to Spall (2000), the user should manually adjust them in order to obtain the best performance of the algorithm. A good performance was obtained by choosing the following values:

$$A=0.430\kappa^{1.661}, \quad \alpha=0.602, \quad c=1, \gamma=0.101,$$

where $\kappa \in (K/2, K)$.

At each iteration t we simulate $2M$ K -dimensional weights from $f(\cdot; \boldsymbol{\delta}^{(t-1)})$, which is either a Dirichlet or a generalized Dirichlet distribution. Next, for each $m=1, \dots, 2M$, the observed log-likelihood is computed, which is the most computationally intensive part of our method. Clearly, as M increases, the noise of the objective function reduces. As noise is lower the convergence is faster. Nevertheless, the computing time per iteration is increased. Consequently, M should be specified as a trade off between the computing time until convergence and computing time per iteration. A typical range for the number of simulated $\boldsymbol{\theta}$ values for our examples is $M \in \{4, 8, 12\}$. Finally, the parameters that control the convergence rule are chosen as: $s=50$ and $v=6$.

As will be demonstrated, the number of mixture components (K) in real RNA-seq datasets can be very large, a typical range is between $(10^3, 10^5)$. In such cases, the parameterization (17) suffers from the drawback that it induces an optimization in a very high dimensional space, making the stochastic approximation algorithm quite slow. For this reason, when the number of components grows large, a simplification is made. In particular, a reduced number of δ_k parameters, say K^* , with $K^* \ll K$, was used. The allocation of the parameters to the generalized Dirichlet distribution is made in the following way:

$$(\tilde{\gamma}_k, \tilde{\gamma}_k^+) = (e^{\delta_{h(k)}} \gamma_k, e^{\delta_{h(k)}} \gamma_k^+), \quad k=1, \dots, K-1, \quad (21)$$

where $h: \{1, \dots, K-1\} \rightarrow \{1, \dots, K^*\}$. The mapping h which assigns parameter $\delta_{h(k)}$ to the slots k and $2k$, $k=1, \dots, K-1$, of the generalized Dirichlet parameters is done by a K -means clustering of the standard VB means. This heuristic reduces the number of parameters and speeds up the convergence of the stochastic approximation algorithm under a small loss in the final marginal likelihood bound. In the following simulated and real data,

whenever the number of components is $K \leq 100$ we set $K^* = K$ (that is, no reduction), while in any other case we used $K^* = 200$ variational parameters.

3 Applications

We illustrate our new algorithm on simulated datasets mimicking the problem of aligning short nucleotide sequences to a reference sequence. Next, the method is applied to six real RNA-seq datasets. Finally, a simulation study under a differential expression setup is provided.

3.1 Simulated data

Let e_j denote an exon sequence, having length equal to ℓ_j , $j=1, \dots, J$. Assume that all exon sequences are different from each other. Consider K discrete sets $\{I_k; k=1, \dots, K\}$, arising by joining distinct combinations of e_j , one after the other, to create transcripts. Let $x_i \sim \sum_{k=1}^K \theta_k \mathcal{U}_{I_k}$, $i=1, \dots, n$ be n randomly sampled short sequences of r consecutive letters from a mixture of uniform distributions defined in I_k , $k=1, \dots, K$. A randomly drawn sequence (x_i) of letters, will potentially map to many sets, that is, there will be more than one k such that: $x_i \in I_k$. Hence, the problem is to estimate the abundance of the generative sets I_k , given the sample.

In total, five datasets (denoted by “A,” “B,” “C,” “D” and “E”) were generated, using the parameter setting shown at Table 1. For dataset A and B the sets I_k , $k=1, 2, 3$ was set as: $I_1=\{e_1, e_3, e_4\}$, $I_2=\{e_1, e_2, e_4\}$, $I_3=\{e_1, e_2, e_3, e_4\}$ and $I_1=\{e_1, e_2\}$, $I_2=\{e_2, e_4\}$, $I_3=\{e_2, e_3, e_4\}$, respectively. The lengths e_j , $j=1, \dots, 4$ were set as (500, 5, 5, 500) and (1000, 500, 5, 1000). For dataset C, $J=20$ sequences was combined to produce a mixture of $K=9$ splices, while for datasets D and E the number of components is equal to 100 and 2000, respectively. The read length was kept constant to $r=50$, except for datasets D and E where a smaller value was chosen in order to account for an increased uncertainty. The range of the true values of the weights was $(10^{-8}, 2 \times 10^{-1})$.

After imposing a $\mathcal{D}(\alpha_1, \dots, \alpha_K)$ prior on θ , with $\alpha_k=1$ for all $k=1, \dots, K$, we applied the three VB algorithms. The estimates of the lower bound of $\log m(\mathbf{x})$ are shown in Table 1. In order to compare these values with the “true” value of the log-marginal likelihood, the method of Chib (1995) was implemented. This method works reasonably well in the low dimensional setting, but it is not efficient to real RNA-seq data where the number of transcripts is much higher. Note here that in this method the computation of $\log \sum_{t=1}^M f(\theta^* | \mathbf{x}, \mathbf{z}^{(t)})$ at a specific high-density point θ^* is involved (where M denotes the number of MCMC iterations). This particular evaluation is very likely to produce numerical errors, hence it is not recommended for large scale datasets (actually, the last row of Table 1 should be interpreted as an upper bound of the log-marginal likelihood, due to numerical underflows in computation of the log of the sum). The deterministic lower bound $\mathcal{L}_1$ arising from the standard VB approximation by optimising (6) is always quite far from the log-marginal likelihood. Next, in row $\mathcal{L}_2(VB)$ notice how the same approximating distribution can be used in order to provide a tighter lower bound. Finally, the proposed optimization technique in $\mathcal{F}_\theta$ and $\mathcal{F}_{\theta^*}$ leads to the improved approximations $\mathcal{L}_2(D)$ and $\mathcal{L}_2(GD)$, respectively.

Figure 1 displays the estimated posterior marginal densities for datasets A, B and C. Long MCMC runs were used to compare the approximation to the ground truth. The MCMC sampler ran for 200,000 iterations, after discarding the first 20,000 as burn-in period. Every 20th iteration was stored in order to compute the density estimate (black line). As previously described, the approximation of the standard VB method is poor in estimating the variance, however in all cases the estimate of the posterior mean is quite accurate. On the other hand, the estimates are getting better according to the Dirichlet (D) or generalized Dirichlet (GD) VB modification. In particular, for dataset A we conclude that both D and GD modifications give almost the same approximation. Moreover, notice that the corresponding bounds at Table 1 are actually the same. This is not true for the other datasets, where we conclude that the (D) modification is better than standard VB, however

Table 1 Parameter setup and log-marginal likelihood bounds for the simulated data: $\mathcal{L}_1$ corresponds to the deterministic standard VB bound arising from Equation (6), $\mathcal{L}_2$ bounds according to standard VB ($\mathcal{L}_2(VB)$), Dirichlet ($\mathcal{L}_2(D)$) and generalized Dirichlet ($\mathcal{L}_2(GD)$) modifications. The log-marginal likelihood estimate (Chib, 1995) is shown at the last row. The estimated standard error (for 100 different runs) for each of the stochastic methods is shown in parenthesis.

Dataset	A	B	C	D	E
K	3	3	9	100	2000
n	2000	2000	5000	20,000	100,000
J	4	4	20	200	200
r	50	50	50	20	20
$\mathcal{L}_1$	-17645.91	-17738.62	-47630.07	-273694.75	-2293796
$\mathcal{L}_2(VB)$	-13827.65 (0.020)	-15340.59 (0.015)	-33338.85 (0.050)	-217439.0 (0.116)	-1099616 (0.763)
$\mathcal{L}_2(D)$	-13825.57 (0.002)	-15340.59(0.015)	-33332.57 (0.016)	-217357.9 (0.198)	-1097032 (0.444)
$\mathcal{L}_2(GD)$	-13825.57(0.001)	-15339.38(0.001)	-33331.83(0.001)	-217350.9(0.174)	-1096977(0.361)
$\log \hat{m}(x)$	-13825.57 (0.011)	-15339.38 (0.010)	-33331.77 (0.137)	-217015.9 (5.280)	-1070386 (80.183)

the problem of variance underestimation is still apparent (notice $k=2,3$ of dataset B and $k=1$ of dataset C at Figure 1). On the other hand, the GD distribution is quite close to the one estimated by MCMC.

These points are strengthened even more in the challenging cases of datasets D and E. In order to compare MCMC and VB we now turn to the coefficient of variation (CV), due to the wide range of the posterior means. The first row of Figure 2 displays the density of CV ratios for the three VB methods over the MCMC estimate. Regarding the standard VB, the ratio is strongly biased towards small values (green curve) indicating CV underestimation. This is corrected when using the Dirichlet modification (red curve), although the variance of the ratio is still high. Finally, the black curve supports the superiority of the proposed GD solution.

Figure 2 Kernel density estimate for the coefficient of variation (CV) ratios: $r_j := CV_k^{(j)} / CV_k^{mcmc}$ of the marginal posterior distribution of $\theta_k | \mathbf{x}$, $k=1, \dots, K$. Simulated datasets D and E (first row) and real RNA-seq datasets (second and third row). The numbers at the legends correspond to $\mathbb{E}(1-r_j)^2$ and $\mathbb{E}(1-r_j)$, with j denoting: standard VB (VB), Dirichlet (D) and generalized Dirichlet (GD).

Table 2 Number of transcripts, sample size and log-marginal likelihood bounds for the RNA-seq datasets.

Data and Accession	Reference	<i>K</i>	<i>n</i>	VB	D	GD
SRR036966 GSM343511	ENSG00000102078	14	61875	−1367478.2	−1367475.8	−1367474.5
SRR036966 GSM343511	ENSG0000009830	24	69920	−1506294.4	−1506293.4	−1506288.8
SRR031708 GSM461176	<i>D. melanogaster</i>	37684	3838411	−83958306.3	−83958266.7	−83957908.6
SRR034782 GSM499403	<i>D. pseudoobscura</i>	103621	2243375	−40040211.10	−40035070.0	−40028979.7
SRR207090 GSM723767	mouse	92350	6250292	−118457087.5	−118455461.5	−118453923.1
SRR032128 GSM432610	pantro2	41489	5794003	−91070395.9	−91069850	−91069087.3

3.2 RNA-seq datasets

The six datasets shown in Table 2 were downloaded from NCBI and mapped to the reference genome (shown at the second column) using Bowtie. Both the number of mixture components (transcripts) and observations (reads) cover a wide range. The method described in Glaus et al. (2012) was implemented in order to compute the likelihood of the *n* reads to the *K* transcripts, as well as to obtain an MCMC sample from the posterior distribution. Finally, the VB methods were applied. The last three columns of Table 2 display the lower bounds of the log-marginal likelihood, where a tighter bound is always estimated as we are moving from left to right.

For the first two small datasets, the marginal posterior densities of the two more highly expressed transcripts are shown in Figure 3. Furthermore, according to the scatterplots on the right, it seems that all VB solutions are able to find the correct correlation between transcripts. However, the covariance is strongly underestimated by standard VB, while the generalized Dirichlet estimate is more accurate. For the last four

Figure 3 Marginal density estimates of the two more highly expressed transcripts for the RNA-seq dataset mapped to ENSG0000009830 (*k*=8, 12) and ENSG00000102078 (*k*=13, 23). MCMC (■), standard VB (---), Dirichlet (· · ·), generalized Dirichlet (— ·). *k*=1, 2, 3 for dataset A, B and *k*=1, 5, 9 for dataset C. The vertical lines denote the estimated MCMC (— ·) and (common) VB means (· · ·). The covariance is shown at the last column.

Figure 4 Kernel density estimate (left) and scatterplots (right) of PPLR for $K=7537$ transcripts according to MCMC, VB and GD.

samples, the ratio of the coefficient of variation of VB methods versus MCMC is displayed in Figure 2. Once again, the underestimation of the (rescaled) standard deviations by the standard VB method and the correction according to generalized Dirichlet distribution is evident.

3.3 Application to differential expression

A by-product of our implementation is to call for differentially expressed transcripts between different samples, based on the approximate distribution arising from VB methods. For this purpose, we simulated a dataset containing $K=7538$ transcript and $n=484,779$ reads. There were two conditions with two replicates and transcript expression was estimated for each case using MCMC, standard VB and the generalized Dirichlet method. In order to compare the VB results with MCMC, the probability of positive log-ratio (PPLR) was used (Liu et al., 2006). Compared to the black curve which is considered as the ground truth, the generalized Dirichlet distribution leads to better performance than standard VB, as shown at Figure 4.

4 Discussion

Variational Bayesian methods for approximating the posterior distribution of the weights of a mixture were presented. Standard VB practice leads to decent performance in terms of posterior means estimation, however the variance is underestimated. A new variational scheme was proposed, by integrating the latent variables out of the VB approximation, while exploiting the means of the standard VB solution. This approach relaxes the assumption of independence (4) between the missing data and the weights of the mixture model, which is taken into account in standard VB. We proved that this leads to a tighter bound of the marginal likelihood. The proposed solution belongs to the richer family of Generalized Dirichlet distributions. Finally, the problem of variance underestimation is corrected. The method was applied to transcript expression estimation with encouraging results.

Regarding the computing time of the proposed methodology, the main computational burden is the evaluation of the noisy bound in (19). In order to provide some insight into this aspect, 50 datasets consisting of $K=10,000$ mixture components were simulated with varying sample size and average number of mapping components per read. The computing time (all coding is done in C++) needed for the convergence of standard VB and generalized Dirichlet VB are displayed in Figure 5. Note here that there is nothing in common

Figure 5 Computing time for standard variational Bayes (VB) and generalized Dirichlet (GD) as a function of sample size (n) and average number of alignments per read, for $K=10,000$ mixture components.

regarding the stopping criterion of each algorithm, so Figure 5 should not be interpreted as a “fair” comparison between these methods. However, it is clear that the improved performance of the new method induces a higher computational cost.

We should stress the fact that in model (1), all component specific densities $f_k(x)$, $k=1,\dots,K$ are known. This is not a general characteristic of mixture posterior distributions, where in such a case all $K!$ permutations of the parameter vector result to the same likelihood. In case of non-identifiable mixture components, an extra effort should be provided, in order to take into account the underlying symmetry of the posterior distribution (Neal, 1999; Frühwirth-Schnatter, 2004). This does not apply in our setup, that is, no “label switching” (see, e.g., Jasra et al., 2005) takes place. In other words, the posterior distribution has exactly one (main) mode. So, the approximation of a unimodal posterior distribution by a unimodal family of distributions (such as the Dirichlet or generalized Dirichlet) is justified, at least from a main-mode hunting point of view. Of course, there is always the possibility of minor modes in the posterior distribution. In this case it is clear that the VB approximations will not result to good performance. However, in all of our simulations with real and synthetic data, no evidence of minor modes was found.

Our current study focused on the correction of the posterior distribution approximation, underlining the improvement of the posterior variance estimate. We didn’t evaluate the impact of our findings on model selection issues, such as comparing different transcriptome definitions on the same data. However, preliminary results (not reported here) indicate that the standard and new way of computing the marginal likelihood bound do not always give the same answer. Our future research will investigate this issue.

Acknowledgements: P. Papastamoulis and M. Rattray were supported by BBSRC award “BB/J009415/1” and EU FP7 project “RADIANT” (grant 305626). J. Hensman was supported by BBSRC award “BB/H018123/2”. P. Glaus was supported by the Engineering and Physical Sciences Research Council “EP/P505208/1”.

References

- Beal, M. and Z. Ghahramani (2003): “The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures,” *Bayesian Statistics*, 7, 453–463.
- Chib, S. (1995): “Marginal likelihood from the Gibbs output,” *Journal of the American Statistical Association*, 90, 1313–1321.
- Corduneanu, A. and C. M. Bishop (2001): Variational bayesian model selection for mixture distributions. In: Jaakkola, T., Richardson, T. (Eds.), *Artificial Intelligence and Statistics*, 2001, pp. 27–34.

- Dempster, J. P., N. M. Laird, and D. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society B*, 39, 1–38.
- Frühwirth-Schnatter, S. (2004): "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques," *The Econometrics Journal*, 7, 143–167.
- Frühwirth-Schnatter, S. (2006): *Finite mixture and Markov switching models*, New York: Springer Series in Statistics, Springer.
- Gelfand, A. and A. Smith (1990): "Sampling-based approaches to calculating marginal densities," *Journal of American Statistical Association*, 85, 398–409.
- Glaus, P., A. Honkela, and M. Rattray (2012): "Identifying differentially expressed transcripts from RNA-Seq data with biological variation," *Bioinformatics*, 28, 1721–1728.
- Hensman, J., P. Glaus, A. Honkela, and M. Rattray (2013): *Fast Approximate Inference of Transcript Expression Levels from RNA-seq Data*. arXiv, URL <http://arxiv.org/abs/1308.5953>.
- Hensman, J., M. Rattray, and N. D. Lawrence (2012) Fast variational inference in the conjugate exponential family. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K., (Eds.), *Advances in Neural Information Processing Systems 25*, pp. 2897–2905, URL http://books.nips.cc/papers/files/nips25/NIPS2012_1314.pdf.
- Hutchison, D. and J. Spall (2004): Stochastic approximation in finite samples using surrogate processes. *Proc. 43rd IEEE Conference on Decision and Control*.
- Jasra, A., C. Holmes, and D. Stephens (2005): "Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, 20, 50–67.
- Jordan, M., Z. Ghahramani, T. Jaakola, and L. Saul (1999): "An introduction to variational methods for graphical models," *Machine Learning*, 37, 183–233.
- Katz, Y., E. Wang, E. Airolidi, and C. Burge (2010): "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nat Methods*, 7, 1009–1015.
- Langmead, B., C. Trapnell, M. Pop, and S. Salzberg (2009): "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, 10, R25.
- Li, B., V. Ruotti, R. Stewart, J. Thomson, and D. CN. (2010): "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, 26, 493–500.
- Liu, X., M. Milo, N. Lawrence, and M. Rattray (2006): "Probe-level measurement error improves accuracy in detecting differential gene expression," *Bioinformatics*, 22, 2107–2113.
- Marin, J., K. Mengerson, and C. Robert (2005): "Bayesian modelling and inference on mixtures of distributions," *Handbook of Statistics*, 25, 577–590.
- Marioni, J., C. Mason, S. Mane, M. Stephens, and Y. Gilad (2008): "RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, 18, 1509–1517.
- McGrory, C. and D. Titterton (2007): "Variational approximations in Bayesian model selection for finite mixture distributions," *Computational Statistics and Data Analysis*, 51, 5352–5367.
- McLachlan, J. and D. Peel (2000): *Finite mixture models*, New York: Wiley.
- Mortazavi, A., B. Williams, K. McCue, L. Schaeffer, and B. Wold (2008): "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat Methods*, 5, 621–628.
- Nariai, N., O. Hirose, K. Kojima, and M. Nagasaki (2013): "TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-seq data by variational Bayesian inference," *Bioinformatics*, 29, 2292–2299.
- Neal, R. (1999): "Erroneous results in "Marginal likelihood from the Gibbs output", Technical report, University of Toronto.
- Spall, J. (1992): "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, 37, 332–341.
- Spall, J. (2000): "Adaptive stochastic approximation by the simultaneous perturbation method," *IEEE Transactions on Automatic Control*, 45, 1839–1853.
- Tanner, M. and W. Wong (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.
- Trapnell, C., L. Pachter, and S. Salzberg (2009): "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, 25, 1105–1111.
- Turro, E., S. Su, A. Goncalves, L. Coin, S. Richardson, and A. Lewin (2011): "Haplotype and isoform specific expression estimation using multi-mapping RNA-Seq reads," *Genome Biology*, 12, R13.
- Wang, Z., M. Gerstein, and M. Snyder (2009): "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet.*, 10, 57–63.
- Wong, T. (1998): "Generalized Dirichlet distribution in Bayesian analysis," *Applied Mathematics and Computation*, 97, 165–181.
- Wong, T. (2010): "Parameter estimation for generalized Dirichlet distributions from the sample estimates of the first and the second moments of random variables," *Computational Statistics and Data Analysis*, 54, 1756–1765.