Statistical mechanics approach to the sample deconvolution problem

N. Riedel[*] and J. Berg[†]

Institut für Theoretische Physik, University of Cologne - Zülpicher Straße 77, 50937 Köln, Germany Sybacol, University of Cologne, Germany

(Received 22 January 2013; published 18 April 2013)

In a multicellular organism different cell types express a gene in different amounts. Samples from which gene expression levels can be measured typically contain a mixture of different cell types; the resulting measurements thus give only averages over the different cell types present. Based on fluctuations in the mixture proportions from sample to sample it is in principle possible to reconstruct the underlying expression levels of each cell type: to deconvolute the sample. We use a statistical mechanics approach to the problem of deconvoluting such partial concentrations from mixed samples, explore this approach using Markov chain Monte Carlo simulations, and give analytical results for when and how well samples can be unmixed.

DOI: 10.1103/PhysRevE.87.042715 PACS number(s): 87.10.Vg, 02.30.Zz

I. INTRODUCTION

Organs in higher organisms are complex tissues containing a variety of different cell types. Brain tissue, for instance, contains not only neurons, but also supporting cells like the astrocytes and oligodendrocytes. Kidney tissue contains the filtering units (podocytes) as well as cells of the capillary system (tubules). Whereas two cells of different cell type have largely the same DNA sequence, only a cell-type specific set of genes will be expressed in a cell [1,2].

Over the last two decades, experimental methods have been developed which allow one to measure the amount of messenger-RNA (mRNA) from different genes in a sample [3,4]. However, one may not be interested in expression levels averaged over all cell types in such a sample, but may want to know the mRNA levels present in the different cell types. A particularly pressing example arises in cancer research, where tissue samples typically contain solid tumor and healthy tissue in unknown proportions [5].

Denoting the proportion of cell type a in sample μ by p_a^μ, and the concentration of mRNA from gene i in cell type a by x_i^a, the concentration of mRNA from gene i in sample μ, X_i^μ is given by

$$X_i^\mu = \sum_{a=1}^{n} p_a^\mu x_i^a + \xi_i^\mu, \tag{1}$$

where the residuals ξ_i^μ stem from sample-specific fluctuations of concentrations or random experimental errors. The number of different cell types is denoted by n. Additionally, we have the constraints $0 < x_i^a, 0 < p_a^\mu < 1$, and $\sum_a p_a^\mu = 1 \; \forall \; \mu$.

Sample deconvolution [6] is the inverse problem of reconstructing the concentrations of gene products x_i^a in each cell type, as well as the mixing proportions p_a^μ of the samples from measurements of mixed samples X_i^μ. The information necessary for this reconstruction must come from fluctuations in the mixing proportions across samples: A cell type with high concentrations of mRNA of genes i and j will induce positively correlated fluctuations of the measurements X_i^μ and X_j^μ as the fraction of this cell type varies from sample to sample (see Fig. 1).

Equation (1) also arises in a broad range of contexts outside molecular biology. It is the fundamental equation of *factor analysis*, a statistical approach where different unknown factors x_a contribute with linear weights p_a to some outcome X [7]. Applications arise, for example, in the context of face recognition [8], data analysis in ecology [9], or fluorescence microscopy [10,11].

There are two questions we address in this paper. First, what are the conditions such that learning from fluctuations is possible at all? And second, how accurately can the reconstruction be made? We first discuss a general constraint on the minimal number of samples needed from linear algebra. We then formulate a Bayesian model for sample deconvolution and study this model from the point of view of statistical physics. We explore the model numerically using Markov chain Monte Carlo (MCMC) sampling of the posterior. Finally, we derive analytical results for the accuracy of reconstruction for a simple, nontrivial case of the problem.

As an initial remark, we derive a constraint on the minimal number of samples needed for reconstruction from linear algebra. Looking at the number of variables in Eq. (1) we have, neglecting the residuals, the matrix equation

$$\underbrace{\begin{pmatrix} X_1^1 & \cdots \\ \vdots & \ddots \\ & \end{pmatrix}}_{MN} = \underbrace{\begin{pmatrix} p_1^1 & \cdots \\ \vdots & \ddots \\ & \end{pmatrix}}_{\geqslant M(n-1)+} \times \underbrace{\begin{pmatrix} x_1^1 & \cdots \\ \vdots & \ddots \end{pmatrix}}_{nN.}, \tag{2}$$

Denoting the total number of samples by M, the number of genes by N and the number of cell types by n, there are MN measurements on the left and $M(n-1) + nN$ unknown variables on the right-hand side. If the number of unknown variables exceeds the number of data points, the system of equations is underdetermined. For a measurement of gene expression levels, the number N of genes will typically be of the order of hundreds or even thousands, exceeding by far the number n of cell types in a sample, or the number M of samples. For $N \gg (n,M)$ the condition not to have an underdetermined set of equations reduces to $M > n$, so the number of samples taken has to be at least larger than the number of cell types.

[*] nriedel@thp.uni-koeln.de

[†] berg@thp.uni-koeln.de

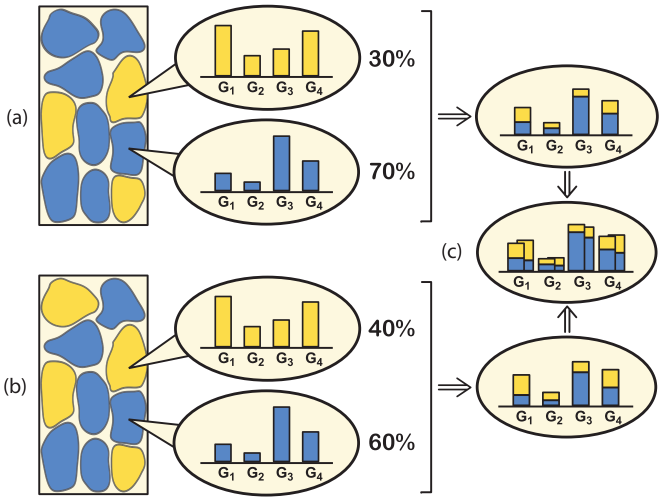

FIG. 1. (Color online) The sample deconvolution problem. (a) A tissue sample containing a mixture of different cell types is taken. The concentration of a particular gene product in the sample is a linear combination of the concentrations in each cell type. (b) Two different tissue samples generally contain different mixtures of cell types, while the concentration of gene products is largely constant across cells of a given type. (c) Part of the variation of expression levels across samples is due to fluctuations in the mixing proportions. This provides the basis for reconstructing concentrations in each cell type and mixing proportions.

II. BAYESIAN MODEL FOR SAMPLE DECONVOLUTION

We formulate a Bayesian approach to sample deconvolution. This approach follows the lines of Bayesian non-negative matrix factorization [12–14], but deviates from previous approaches to sample deconvolution [15–17]. For concreteness we assume that the residuals ξ_i^μ in Eq. (1) are independent and identically distributed Gaussian variables. Other distributions will be discussed below. These residuals stem from fluctuations in the concentrations of the same cell type ("biological noise") and random experimental error ("technical noise"). Given the mixing proportions p_a^μ and the concentrations in cell types x_i^a, the distribution of residuals induces the distribution of the measurements X_i^μ

$$P(\mathbf{X}|\mathbf{p},\mathbf{x}) = \left(2\pi\sigma_\xi^2\right)^{-\frac{MN}{2}} e^{-\mathcal{H}(\mathbf{p},\mathbf{x};\mathbf{X})}, \tag{3}$$

with the Hamiltonian

$$\mathcal{H}(\mathbf{p},\mathbf{x};\mathbf{X}) = \sum_{\mu,i} \frac{\left(X_i^\mu - \sum_a p_a^\mu x_i^a\right)^2}{2\sigma_\xi^2}. \tag{4}$$

Bold symbols are used to denote matrices: $(\mathbf{X})_i^\mu = X_i^\mu$, etc.

The quantity of interest; namely, the probability of a given set of mixtures and of concentrations in cell types given the measurements, follows from Bayes theorem; the so-called posterior probability $P(\mathbf{p},\mathbf{x}|\mathbf{X}) = \frac{P(\mathbf{p},\mathbf{x})P(\mathbf{X}|\mathbf{p},\mathbf{x})}{P(\mathbf{X})}$ is expressed in terms of the likelihood (3), the prior $P(\mathbf{p},\mathbf{x})$, and the marginal likelihood $P(\mathbf{X})$ (which for a given set of measurements is just a multiplicative constant).

For the prior $P(\mathbf{p},\mathbf{x})$, a particular choice has to be made. A natural choice for the mixing proportions is the Dirichlet distribution, which restricts the possible points to the simplex and automatically fulfills the condition $\sum_a p_a^\mu = 1$. We denote the parameters of the Dirichlet distribution by α_a. For the expression levels we choose independent Gamma distributions with shape parameters k_a and scale parameters θ_a, allowing for different parameters of this distribution for each cell type. The Gamma distribution has positive support only, and is used widely to describe the distribution of gene expression levels [18]. In the general algorithm we discuss below, any distribution can be implemented. For this particular choice of the prior we get

$$P(\mathbf{p},\mathbf{x}|\mathbf{X}) = \left[\prod_\mu \delta\left(\sum_{a=1}^n p_a^\mu - 1\right)\right] e^{-\mathcal{H}_{\text{Bayes}}(\mathbf{p},\mathbf{x};\mathbf{X})}/Z_{\mathbf{X}}, \tag{5}$$

with the Hamiltonian

$$\begin{aligned}\mathcal{H}_{\text{Bayes}}(\mathbf{p},\mathbf{x};\mathbf{X}) = {} & \sum_{\mu i} \frac{\left(X_i^\mu - \sum_a p_a^\mu x_i^a\right)^2}{2\sigma_\xi^2} \\ & + \sum_{ai}\left[(k_a - 1)\ln x_i^a - \frac{x_i^a}{\theta_a}\right] \\ & + \sum_{\mu a}(1-\alpha_a)\ln p_a^\mu. \end{aligned} \tag{6}$$

The first term in Eq. (6) comes from the likelihood, penalizing the deviation of a possible solution $\{p_a^\mu, x_i^a\}$ from the measurement $\{X_i^\mu\}$. The second and third term are the Gamma priors for the expression patterns and the Dirichlet density for the mixing proportions, respectively. The values of the mixing proportions are restricted to the simplex using a δ function in Eq. (5). The posterior distribution (5) describes the state of our knowledge of the mixing proportions and concentrations in each cell type, given the measurements. From the perspective of statistical physics, the mixing proportions and concentrations in each cell type define a phase space, and the posterior distribution (5) defines a Hamiltonian (6) describing how strongly the probability measure of mixing proportions and concentrations is focused in particular parts of this phase space. The partition function

$$Z_{\mathbf{X}}(\beta) = P(\mathbf{X}) = \mathrm{Tr}_{\mathbf{p},\mathbf{x}}\, e^{-\beta\mathcal{H}_{\text{Bayes}}(\mathbf{p},\mathbf{x};\mathbf{X})}, \tag{7}$$

with $\mathrm{Tr}_{\mathbf{p},\mathbf{x}} = \int d\mathbf{p} \int d\mathbf{x} \prod_\mu \delta(\sum_{a=1}^n p_a^\mu - 1)$ sets out the statistical mechanics of sample deconvolution at $\beta = 1$. The corresponding entropy $S = \frac{\partial}{\partial\beta}\frac{1}{\beta}\ln Z_{\mathbf{X}}|_{\beta=1}$ is a measure of the uncertainty of reconstruction.

III. SAMPLE DECONVOLUTION ALGORITHM

First, we explore the Boltzmann posterior distribution (5) numerically. For this purpose we use MCMC sampling of the posterior as has been done before for similar Bayesian models in the context of non-negative matrix factorization (NMF) [12–14]. The MCMC sampling explores the entire space of reconstructions weighted with the posterior probability. An arbitrary starting configuration $\mathcal{C}_0 = \{\mathbf{p},\mathbf{x}\}$ is chosen and from this starting point a new neighboring configuration $\mathcal{C}_1$ is generated by randomly increasing or decreasing one of the free parameters by a small amount within the positivity constraints and the constraint on the mixing proportions.

The new configuration is accepted with the probability $p_{acc} = \min(1, e^{-(\mathcal{H}_1 - \mathcal{H}_0)})$ with energies $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding to configuration $\mathcal{C}_0$ and $\mathcal{C}_1$, respectively (Metropolis rule). Since only the ratio of the posterior probability of two configurations enters, the marginal likelihood $P(\mathbf{X})$ is not needed.

We test this algorithm on artificially generated datasets. For this purpose a target solution $\mathbf{x}^T$ is drawn randomly from Gamma distributions with shape parameters k_a and scale parameters θ_a. The target mixture $\mathbf{p}^T$ is drawn randomly from a Dirichlet distribution with parameters α_a. Then the measurements $\mathbf{X}^T$ are generated according to Eq. (1) by adding Gaussian noise of variance $\sigma_\xi^{T\,2}$. The Metropolis rule is used to sample the posterior (5) of the mixtures p_a^μ, the concentrations x_i^a, and the parameters describing the prior $\bar{x}_a$, $\sigma_{x,a}^2$, α_a, and σ_ξ^2. In this way only the general shape of the prior distribution needs to be chosen, the actual parameters of the distribution are estimated from the data.

This MCMC sampling allows one to sample the regions in phase space where the posterior probability is high. An estimate of the mixtures and concentrations is provided by the mean values of x_i^a and p_a^μ obtained by averaging over a large number of configurations visited during the sampling process. In addition to this posterior mean, we also compute the standard deviations of x_i^a and p_a^μ under the posterior. These standard deviations quantify the remaining uncertainty in the reconstruction, given the noise in the limited amount of data, and can serve as error estimates of the reconstruction. In Fig. 2 we show for each variable x_i^a the mean and standard deviation (as error bars) under the posterior (5) against the targets x_i^{Ta}.

The Bayesian approach differs from a class of popular algorithms for sample deconvolution based on non-negative matrix factorization (NMF) [15–17]. NMF aims to invert the relationship $\mathbf{X} = \mathbf{px}$, while keeping all matrix entries positive. Starting with an initial guess of $\mathbf{x}$, the matrix $\mathbf{p}$ is calculated that minimizes the l_2-norm

$$\|\mathbf{X} - \mathbf{p}\cdot\mathbf{x}\|_{\mathrm{F}} \equiv \sqrt{\sum_{\mu,i}\left(X_i^\mu - \sum_a p_a^\mu x_i^a\right)^2}$$

(Frobenius norm). The minimization proceeds under the additional constraint of non-negative matrix entries. From this estimate of $\mathbf{p}$ an improved estimate for $\mathbf{x}$ is obtained by applying the procedure in turn to $\mathbf{x}$. Iterating these steps the distance $\|\mathbf{X} - \mathbf{px}\|_{\mathrm{F}}$ of the reconstructed solution from the data matrix never increases. This implies a Gaussian noise model and corresponds to a minimization of $\mathcal{H}_L$ in Eq. (4). Convergence of the iterative scheme is thus guaranteed and reaches a local minimum of the Hamiltonian (4).

As an example from the class of NMF algorithms we applied the *deconf* algorithm [15] to the same artificial data as used for the Bayesian algorithm. Figure 2 shows how the Bayesian approach outperforms the NMF-based *deconf*. Thirty samples are needed by *deconf* to reach a reconstruction accuracy similar to the Bayesian algorithm using only five samples. Another NMF algorithm we tried [17] achieved an even lower accuracy.

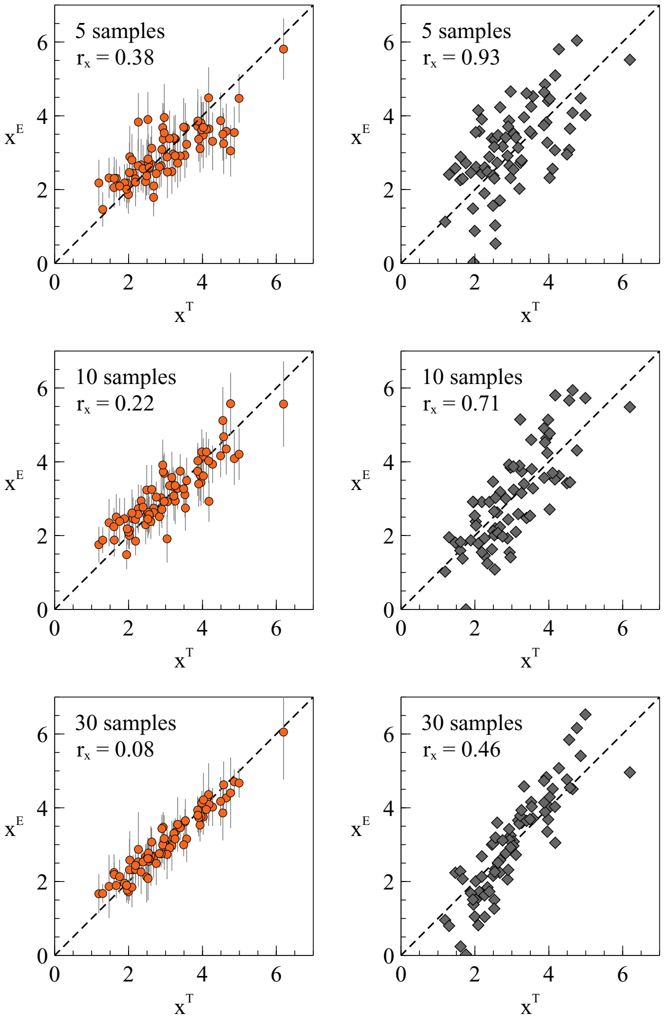

FIG. 2. (Color online) Reconstruction by Bayesian algorithm (left) and the *deconf* algorithm (right) for different number of samples. The reconstruction estimates $\mathbf{x}^E$ are obtained from the same target solution $\mathbf{x}^T$ drawn from a Gamma distribution for all plots. The algorithm estimates the values of $\mathbf{x}^E$, $\mathbf{p}^E$ and the prior parameters $\bar{x}_a$, $\sigma_{x,a}^2$, α_a, and σ_ξ^2. The reconstruction with the Bayesian algorithm is clearly more accurate: the reconstruction accuracy $r_x = \frac{1}{nN}\sum_{a,i}(x_i^{Ea} - x_i^{Ta})^2$ of the Bayesian algorithm using 5 samples is comparable to the accuracy of the *deconf* algorithm using 30 samples. Additionally, the standard deviation of the posterior distribution serves as a natural measure for uncertainty of the estimate, giving rise to the individual error bars in the case of the Bayesian algorithm. These error bars provide a measure for the precision of the reconstruction. Parameters used for both algorithms are $N = 500$, $n = 3$, $M = 5$, 10, and 30, $\alpha = 30$ for all cell types equally, $k = 9$ and $\theta = 1/3$ also for all cell types equally, and $\sigma_\xi^T = 0.1$. In order to distinguish single points with their corresponding error bars, the values of only one in twenty of the 500 genes are plotted.

Of course prior information in Eq. (6) on the distribution of targets facilitates the reconstruction for the Bayesian algorithm. However, the prior is not responsible for the entire difference in performance between the Bayesian and the NMF approach. Figure 3 shows that the effect of the prior is largest when the number of samples is small. This is to be expected, since the relative contribution of the prior to Eq. (6) increases when the amount of information coming from the measurements decreases. Even without using prior information, the

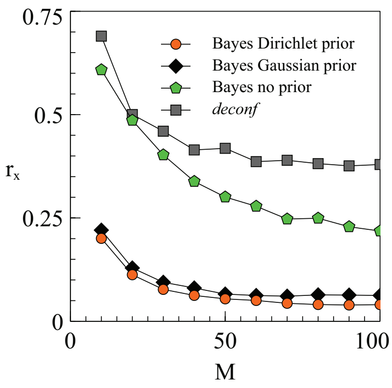

FIG. 3. (Color online) The prior information is most important when few samples are available. We plot the average reconstruction accuracy $r_x = \frac{1}{nN}\sum_{a,i}(x_i^{Ea} - x_i^{Ta})^2$ against the number of samples M for the Bayesian algorithm with different priors and the *deconf* algorithm. For only few samples the Bayesian algorithm with prior (circles) clearly outperforms the NMF (squares) but also the Bayesian algorithm without prior (pentagons). Without the use of prior information the Bayes algorithm is performing equally to the NMF algorithm for a small number of samples. With increasing sample number, the Bayes algorithm without prior learns much faster from the additional information, gradually approaching the performance of the Bayes algorithm with prior. We note that replacing both the Dirichlet prior and the Gamma prior with Gaussian priors (diamonds) for datasets with mixing proportions and expressions levels drawn from the Dirichlet and Gamma distribution, respectively, has very little effect. Simulation parameters are as in Fig. 2.

Bayesian algorithm performs as well as the NMF algorithm in the low sample number regime. For an increasing number of samples the performance of sampling without prior approaches the performance with use of prior information. This is expected as well, since then the relative contribution of the prior to Eq. (6) becomes asymptotically negligible. NMF approaches, formulated as optimization problems, give a point estimate in phase space that reproduces the matrix of measurements $\mathbf{X}$ as closely as possible, leading to the well-known problem of overfitting [19,20].

IV. PARTITION FUNCTION AND THEORETICAL RECONSTRUCTION ACCURACY

We now address the theoretical limit of sample deconvolution. For a given set of measurements $\mathbf{X}$, the statistics of mixtures and concentrations is described by the partition function (7). To keep the analytical calculations tractable, we approximate both the Dirichlet distribution for the mixing proportions and the Gamma distribution for the expression levels by a Gaussian distribution. Especially if the variance of the Dirichlet distribution is small, we expect this to be a good approximation. To validate this approximation, we tried a Gaussian prior on the datasets actually generated by the Dirichlet and Gamma distributions in the numerical simulations (see Fig. 3). For the parameters used here, we observe only small effects from using the approximate Gaussian priors. The Gaussian priors change the Hamiltonian (6) to

$$\mathcal{H}_{\text{Bayes}}(\mathbf{p},\mathbf{x};\mathbf{X}) = \sum_{\mu i} \frac{\left(X_i^\mu - \sum_a p_a^\mu x_i^a\right)^2}{2\sigma_\xi^2} + \sum_{\mu a} \frac{\left(p_a^\mu - \bar{p}_a\right)^2}{2\sigma_{p,a}^2} + \sum_{ai} \frac{\left(x_i^a - \bar{x}_a\right)^2}{2\sigma_{x,a}^2}, \qquad (8)$$

with means $\bar{p}_a/\bar{x}_a$ and variances $\sigma_{p,a}^2/\sigma_{x,a}^2$ as new parameters for the distributions of mixing proportions and expression levels, respectively. The partition function $Z_{\mathbf{X}}$ in Eq. (7) is now defined with the new Hamiltonian (8).

Suppose now that a given set of measurements is generated using Eq. (3) from underlying mixing proportions $\mathbf{p}^T$ and concentrations $\mathbf{x}^T$. These are the targets that the reconstruction aims for. In order to explore the behavior of the system for typical realizations of these targets, the quenched average of $\ln Z_{\mathbf{X}}$ over $\mathbf{p}^T$ and $\mathbf{x}^T$ needs to be computed [21]. We restrict ourselves to the so-called annealed approximation and calculate the average of $Z_{\mathbf{X}}$ over the target mixtures and concentrations. We will show numerically that the difference between quenched and annealed results are small for a reasonable choice of parameters. The annealed average over $Z_{\mathbf{X}}$ is given by

$$\langle\langle Z_{\mathbf{X}}\rangle\rangle = \int d\mathbf{p}^T \int d\mathbf{x}^T \int d\boldsymbol{\xi} \int d\mathbf{X}\, P(\mathbf{p}^T,\mathbf{x}^T,\boldsymbol{\xi}) \times \left[\prod_\mu \delta\left(\sum_{a=1}^n p_{\mu a}^T - 1\right)\right] \delta(\mathbf{X} - \mathbf{p}^T\mathbf{x}^T - \boldsymbol{\xi}) Z_{\mathbf{X}}. \qquad (9)$$

The average $\langle\langle\cdot\rangle\rangle$ is over all data matrices weighted by their probabilities under the generative model $P(\mathbf{p}^T,\mathbf{x}^T,\boldsymbol{\xi}) = \exp\{-\frac{1}{2\sigma_\xi^2}\boldsymbol{\xi}^2 - \frac{1}{2\sigma_x^2}(\mathbf{x}^{\mathbf{T}} - \bar{\mathbf{x}})^2 - \frac{1}{2\sigma_p^2}(\mathbf{p}^{\mathbf{T}} - \bar{\mathbf{p}})^2\}$ for given parameters $\bar{x}_a/\bar{p}_a$, $\sigma_{x,a}^2/\sigma_{p,a}^2$, and σ_ξ^2. This average of the partition function $Z_{\mathbf{X}}$ leads to a large set of integrals, which in the thermodynamic limit $N \to \infty$ can be evaluated using the saddle-point approximation [21–24]. For the thermodynamic limit we consider the scaling ansatz $M = \alpha N$ and $\sigma_\xi^2 = \tilde{\sigma}_\xi^2 N$. In a concrete situation N is of course finite, and α, $\tilde{\sigma}_\xi$ will be small (but also finite). In a lengthy but standard calculation we obtain the saddle point equations

$$\begin{aligned} \hat{q}_{ab} &= \frac{\alpha}{4\tilde{\sigma}_\xi^2}\langle p_a p_b\rangle_{\mathbf{p},\mathbf{p}^T}, \quad & q_{ab} &= \langle x_a x_b\rangle_{\mathbf{x},\mathbf{x}^T}, \\ \hat{q}_{ab}^T &= -\frac{\alpha}{2\tilde{\sigma}_\xi^2}\left\langle p_a^T p_b\right\rangle_{\mathbf{p},\mathbf{p}^T}, \quad & q_{ab}^T &= \left\langle x_a^T x_b\right\rangle_{\mathbf{x},\mathbf{x}^T}, \\ \hat{q}_{ab}^{TT} &= \frac{\alpha}{2\tilde{\sigma}_\xi^2}\left\langle p_a^T p_b^T\right\rangle_{\mathbf{p},\mathbf{p}^T}, \quad & q_{ab}^{TT} &= \left\langle x_a^T x_b^T\right\rangle_{\mathbf{x},\mathbf{x}^T}, \end{aligned} \qquad (10)$$

with averages defined as

$$\langle(\cdots)\rangle_{\mathbf{p},\mathbf{p}^T} = \frac{1}{Z_p}\int\prod_a d\,p_a^T \int\prod_a d\,p_a\,(\cdots) \times \delta\left(\sum_{a=1}^n p_a^T - 1\right)\delta\left(\sum_{a=1}^n p_a - 1\right) \times \exp\left\{-\frac{1}{4\tilde{\sigma}_\xi^2}\sum_{ab}\left(p_a p_b q_{ab} + p_a^T p_b^T q_{ab}^{TT}\right.\right.$$

$$- 2 p_a^T p_b q_{ab}^T) - \frac{1}{2\sigma_p^2}\sum_a (p_a - \bar{p})^2 - \frac{1}{2\sigma_p^2}\sum_a \left(p_a^T - \bar{p}\right)^2 \Big\} \tag{11}$$

and

$$\langle\langle \cdots \rangle\rangle_{\mathbf{x},\mathbf{x}^T} = \frac{1}{Z_x}\int \prod_a d\, x_a^T \int \prod_a d\, x_a (\cdots) \times \exp\Big\{ -\sum_{ab} \left(x_a x_b \hat{q}_{ab} + x_a^T x_b \hat{q}_{ab}^T + x_a^T x_b^T \hat{q}_{ab}^{TT}\right) - \frac{1}{2\sigma_x^2}\sum_a (x_a - \bar{x})^2 - \frac{1}{2\sigma_x^2}\sum_a \left(x_a^T - \bar{x}\right)^2 \Big\}. \tag{12}$$

The q variables in Eq. (10) are the order parameters of the system. They are connected to the Euclidean distance between the reconstruction and its target, $r_x = \frac{1}{nN}\sum_{a,i}(x_i^a - x_i^{Ta})^2$ and similarly $r_p = \frac{1}{nM}\sum_{a,\mu}(p_a^\mu - p_a^{T\mu})^2$, through the relationships $r_x = \frac{N}{n}\sum_a(q_{aa} - 2q_{aa}^T + q_{aa}^{TT})$ and $r_p = \frac{N}{n}\sum_a(\hat{q}_{aa} - 2\hat{q}_{aa}^T + \hat{q}_{aa}^{TT})$. The saddle-point equations (10) have to be solved numerically. We note the formal similarity of the number of cell types with replicas used to calculate quenched averages. A quenched calculation of this system would thus result in a system of equations bearing a two-replica structure [25].

Solving the saddle-point equations (10) numerically for the first nontrivial case of $n = 2$ cell types, we are particularly interested in the difference between target and the reconstructed mixtures and concentrations and how this difference depends on the number of samples M. To this end, we define a normalized order parameter $r_x^n \equiv \frac{r_x}{2\sigma_x}$, which is zero for perfect reconstruction and one for random guessing from the prior distribution of the target solution. Figure 4 shows how the reconstruction improves with increasing number of samples M for a high (top) and low (bottom) noise level, respectively. The improvement of the reconstruction is not surprising because each additional sample brings different mixing proportions and induces correlations in the measurements X_i^μ across genes, from which both concentrations in the cell types and mixing proportions can be reconstructed. There is no finite threshold in the number of samples below which reconstruction is impossible, at any nonzero value of $\alpha = M/N$ the reconstruction is better than a random choice of concentrations and mixtures.

To compare these results with numerical simulations, we evaluated the quenched average by drawing target mixtures and concentrations from the Gaussian prior distribution and the measurements X_i^μ generated from Eq. (3). Then Markov chain Monte Carlo (MCMC) sampling is used to draw the reconstructed mixtures and concentrations from the Boltzmann posterior distribution (6). We also simulated the annealed average by including prior for the target mixtures and concentrations into the Hamiltonian and sampling over the target mixtures and concentrations as well. Very good agreement between these numerical and the analytical results in the high noise regime is seen in Fig. 4 (top). In the low noise regime, numerical simulations show a deviation of the annealed approximation from the quenched result in a certain regime of α. This points out the limitations of the annealed approximation. Here, a full, quenched treatment of the problem would be necessary to get an accurate description of the reconstruction accuracy for all parameter regimes.

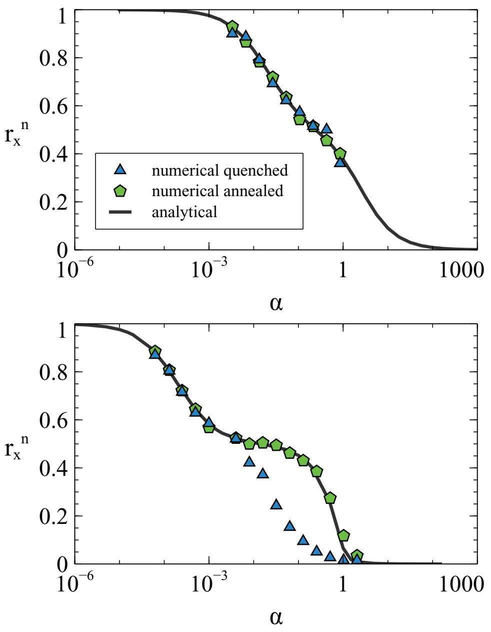

FIG. 4. (Color online) Top: High noise regime, $\sigma_\xi = 2$. Here, numerical simulations corresponding to annealed and quenched average are in good agreement with the analytical annealed result. With the increase of the fraction of samples α the reconstruction accuracy improves from random guessing ($r_x^n = 1$) to a perfect reconstruction ($r_x^n = 0$), but many samples are needed ($\alpha = M/N > 1$) to reach a low r_x^n. Annealed and quenched simulation results are in good agreement, indicating the annealed calculation to be a good approximation to the quenched calculation in this case. Bottom: Low noise regime, $\sigma_\xi = 0.4$. The whole curve for the reconstruction accuracy is shifted to the left; now, fewer samples are needed ($\alpha = M/N < 1$) to obtain a precise reconstruction. In this regime the annealed approximation does not predict the quenched average well for the whole range of α, pointing out the limitations of the annealed approximation in this case. The parameters are chosen in both cases as: $M = 1 - 500$ at $N = 500$, $n = 2$, $\bar{p} = 1/2$, and $\sigma_p = 0.1$ for all cell types equally; $\bar{x} = 3$ and $\sigma_x = 1$ also for all cell types equally.

For the calculation of the partition function (9) we assumed the concrete case of Gaussian distributed residuals. For the algorithm, there is no loss of generality involved, any well-behaved probability density can be used in Eq. (5), leading generally to a Hamiltonian that is not quadratic. For the analytic calculation, Taylor expanding such a Hamiltonian around $\mathbf{X}$ would give $\mathcal{H}(\mathbf{p},\mathbf{x};\mathbf{X}) = a_0 + a_1(\mathbf{X} - \mathbf{px}) + a_2(\mathbf{X} - \mathbf{px})^2 + \cdots$. Our analytical calculation focuses exclusively on the second-order term. The first-order term alone (apart from not being normalizable), would induce no correlations between the targets $\mathbf{p}^T,\mathbf{x}^T$ and the estimated solution $\mathbf{p},\mathbf{x}$. It may be possible, at least in principle, to evaluate the integrals arising from even polynomials beyond second order, but the resulting expressions will not admit simple order parameters. The Gaussian distributed residuals (3) thus define the nontrivial yet tractable model of sample deconvolution.

V. OUTLOOK

In summary, we have developed a Bayesian model for reconstructing cell-type specific gene concentrations from samples containing an unknown mixture of cell types with unknown concentrations. To explore the reconstruction ability of this method we used MCMC sampling of the solution space weighted by the posterior distribution. This turns out to outperform methods based on minimizing the distance between the matrix of measurements $\mathbf{X}$ and the matrix product of mixing proportions $\mathbf{p}$ and concentrations $\mathbf{x}$. Formulating the problem in the language of statistical mechanics, we have obtained an analytical solution for the reconstruction accuracy using the annealed approximation for the specific case of $n = 2$ cell types. This solution can be extended easily for any finite number of cell types. This annealed approximation shows deviations from the quenched result in a certain parameter regime. In order to close this gap, a full, quenched calculation would be necessary.

As ever, the proper choice of the prior may be a delicate step, and Gaussian noise terms or priors need not optimally describe real datasets. A study based on experimental datasets would be needed to settle this issue. But if a better choice for the prior is found, it would be straightforward to implement into the algorithm, leaving the rest of the Bayesian framework unchanged.

ACKNOWLEDGMENTS

This study was funded by the BMBF through Sybacol. We gratefully acknowledge discussions with Roman Müller, Bernhard Schermer, and Thomas Benzing from the Kidney Research Center Cologne.

[1] C. Palmer, M. Diehn, A. Alizadeh, and P. Brown, BMC Genomics **7**, 115 (2006).

[2] D. W. Galbraith and K. Birnbaum, Annu. Rev. Plant Biol. **57**, 451 (2006).

[3] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, Science **270**, 467 (1995).

[4] D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis, Proc. Natl. Acad. Sci. USA **94**, 13057 (1997).

[5] S. Cleator, A. Tsimelzon, A. Ashworth, M. Dowsett, T. Dexter, T. Powles, S. Hilsenbeck, H. Wong, C. Osborne, P. O'Connell, and J. Chang, Breast Cancer Res. Treat. **95**, 229 (2006).

[6] J. Clarke, P. Seo, and B. Clarke, Bioinformatics **26**, 1043 (2010).

[7] D. Barber, *Bayesian Reasoning and Machine Learning* (Cambridge University Press, Cambridge, 2010).

[8] J. Elder, S. Prince, Y. Hou, M. Sizintsev, and E. Olevskiy, Int. J. Comput. Vis. **72**, 47 (2007).

[9] A. H. Hirzel, J. Hausser, D. Chessel, and N. Perrin, Ecology **83**, 2027 (2002).

[10] T. Zimmermann, in *Microscopy Techniques*, Advances in Biochemical Engineering, edited by J. Rietdorf (Springer, Berlin, Heidelberg, 2005), Vol. 95, pp. 245–265.

[11] R. Lansford, G. Bearman, and S. E. Fraser, J. Biomed. Opt. **6**, 311 (2001).

[12] M. Ochs, R. Stoyanova, F. Arias-Mendoza, and T. Brown, J. Magn. Reson. **137**, 161 (1999).

[13] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, IEEE T. Signal Process. **54**, 4133 (2006).

[14] M. N. Schmidt and H. Laurberg, Comput. Intell. Neurosci. **2008**, 361705 (2008).

[15] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G. Black, J. Selbig, S. Parida, S. Kaufmann, and M. Jacobsen, BMC Bioinf. **11**, 27 (2010).

[16] H. Lahdesmaki, l. Shmulevich, V. Dunmire, O. Yli-Harja, and W. Zhang, BMC Bioinf. **6**, 54 (2005).

[17] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini, Bioinformatics **17**, S279 (2001).

[18] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K.-W. Tsui, J. Comput. Biol. **8**, 37 (2001).

[19] M. Opper, *The Handbook of Brain Theory and Neural Networks*, edited by M. A. Arbib (MIT Press, Cambridge, 2003).

[20] A. Engel and C. van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[21] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[22] E. Gardner, J. Phys. A: Math. Gen. **21**, 257 (1988).

[23] E. Gardner and B. Derrida, J. Phys. A: Math. Gen. **21**, 271 (1988).

[24] M. Opper, Phys. Rev. E **51**, 3613 (1995).

[25] R. Monasson, Phys. Rev. Lett. **75**, 2847 (1995).