Bioinformatics Advance Access published March 10, 2014

Probabilistic PCA of censored data: accounting for uncertainties in the visualisation of high-throughput single-cell qPCR data

Florian Buettner [1,*], Victoria Moignard[2], Berthold Göttgens[2] and Fabian J. Theis [1,3]

[1]Institute of Computational Biology, Helmholtz-Zentrum München, 85764 Neuherberg, Germany
[2]University of Cambridge, Department of Haematology, Cambridge Institute for Medical Research and Wellcome Trust & MRC Cambridge Stem Cell Institute, Cambridge, CB2 0XY, United Kingdom
[3]Department of Mathematics, TU München, 85748 Garching, Germany

Associate Editor: Dr. Janet Kelso

ABSTRACT

Motivation: High-throughput single-cell qPCR is a promising technique allowing for new insights in complex cellular processes. However, the PCR reaction can only be detected up to a certain detection limit, while failed reactions could be due to low or absent expression and the true expression level is unknown. As this censoring can occur for high proportions of the data, it is one of the main challenges when dealing with single-cell qPCR data. PCA is an important tool for visualising the structure of high-dimensional data as well as for identifying sub-populations of cells. However, to date it is not clear how to perform a PCA of censored data. We present a probabilistic approach which accounts for the censoring and evaluate it for two typical data-sets containing single-cell qPCR data.

Results: We use the Gaussian Process Latent Variable Model (GPLVM) framework to account for censoring by introducing an appropriate noise model and allowing a different kernel for each dimension. We evaluate this new approach for two typical qPCR data-sets (of mouse embryonic stem cells and blood stem/progenitor cells respectively) by performing linear and non-linear probabilistic PCA. Taking the censoring into account results in a 2D representation of the data which better reflects its known structure: in both data-sets our new approach results in a better separation of known cell types and is able to reveal subpopulations in one data-set which could not be resolved using standard PCA.

Availability: The implementation was based on the existing GPLVM toolbox[1]; extensions for noise models and kernels accounting for censoring are available from http://icb.helmholtz-muenchen.de/censgplvm.

Contact: fbuettner.phys@gmail.com

1 INTRODUCTION

1.1 High-throughput single-cell qPCR

In order to gain fundamental insights into complex cellular processes, it is necessary to observe individual cells. One such process is the transcriptional control of cell fate decisions, where it is crucial to quantify the gene expression of individual cells as cell fate decisions are made on a single-cell level. In contrast to single-cell measurements, conventional experimental techniques measure gene expression from pools of cells masking heterogeneities within cell populations which may be important for understanding underlying biological processes (Guo *et al.*, 2010; Dalerba *et al.*, 2011; Pina *et al.*, 2012; Moignard *et al.*, 2013; Dominguez *et al.*, 2013). Recent technical advances facilitate the simultaneous measurement of tens to thousands of genes in hundreds of individual cells (Taniguchi *et al.*, 2009). As experimental techniques advance and new types of data are generated, it is important to develop sound computational methods that are able to adequately deal with uncertainties inherent in the experimental technique and allow for a comprehensive analysis of these new types of data. Currently, the mRNA content of single cells can be analysed using high-throughput qPCR platforms, such as the Fluidigm BioMark HD or using deep sequencing (RNA-Seq).

In single-cell quantitative real time polymerase chain reaction (qPCR), RNA is extracted from single cells and cDNA is synthesised. This is followed by a pre-amplification step and qPCR detection. In practice, this procedure results in a limit of detection below which gene activity cannot be quantified. Gene expression is typically measured in cycles (Ct) and depending on the analysed cell types and genes, the limit of detection (LOD) Ct value can be defined as a 99% detection probability of the qPCR reaction and typically corresponds to approximately 2-10 mRNA molecules per reaction chamber (Fluidigm Corporation, 2012). This corresponds to a censoring in the sense that for Ct values greater than LOD Ct the true Ct number cannot be established. This censoring typically occurs for a large number of cells (see figure 1 for values of 2 typical data-sets) and is one of the main challenges when dealing with data from single-cell qPCR experiments. For cases in which non-detection corresponds to a lack of transcription, the true Ct value would be infinity (McDavid *et al.*, 2013) whereas for cases in which non-detection corresponds to a non-negligible amount of transcription the true underlying Ct value would be closer to LOD Ct; as the distribution of Ct values extends continuously until the LOD (figure 1 C and D) this suggests that both scenarios can be encountered in practice.

As high-throughput single-cell qPCR is a relatively new technique

*to whom correspondence should be addressed

[1] https://github.com/SheffieldML/GPmat

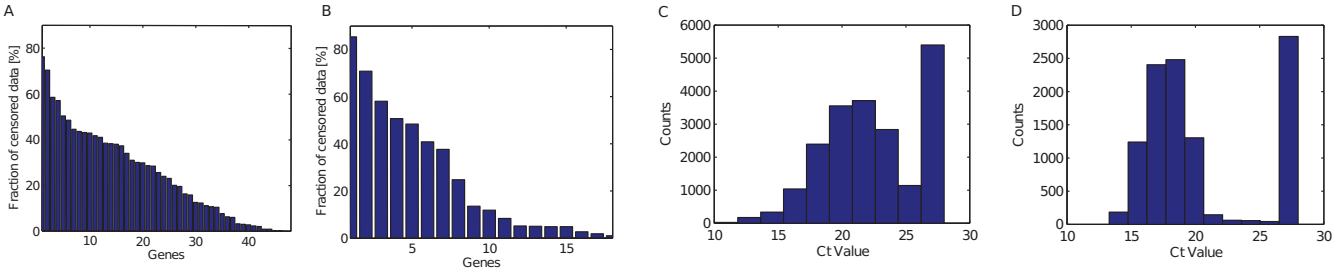


Fig. 1. Fraction of censored data for two typical data sets: A, fraction of non-detects in mESC data data, resovled by genes and B, feactions of non-detects in and blood stem cell data. Genes sorted in descending order of fraction of censored values. C, distribution of Ct values for mESC data and D, blood stem/progenitor cell data. The long tail of high Ct values continues until the LOD.

this issue of censoring has not been addressed systematically and simple work-arounds such as substituting all censored data-points with the LOD Ct value are commonly used (Guo *et al.*, 2010; Dalerba *et al.*, 2011; Pina *et al.*, 2012).

Recently, McDavid *et al.* (2013) have systematically addressed this issues by proposing a customised approach for univariate testing of differential gene expression of single-cell qPCR data which explicitly takes the component of non-detected qPCR reaction into account. While the authors did not address implications of the limit of detection for multivariate analyses such as PCA, this highlights the need for new algorithms addressing statistical and analytical challenges of single-cell qPCR data.

Other sources of uncertainty on a cell-wise level such as effects due to variations in cell size, can be corrected for by measuring a set of housekeeping genes and subtracting the mean expression from the measured Ct number. Similarly, uncertainties can be corrected which occur due to the batch-wise processing of cells on arrays and variations in PCR efficiency between batches.

1.2 PCA of censored data

A common part of multivarate analysis of single-cell qPCR data is principal component analysis (PCA). This allows for a visualisation of the variation in gene expression within and across different cell populations as well as the identification sub-populations in a large group of heterogeneous cells (Guo *et al.*, 2010; Dalerba *et al.*, 2011). Recently we have shown that it is desirable to also apply non-linear generalisations of PCA as this can allow for a better identification of novel sub-populations (Buettner and Theis, 2012). For many statistical methods such as regression, algorithms to deal with censored data have been established. For example, censored values can be substituted, Tobit regression can be performed or data can be deleted, treated as missing or imputed according to some probability distribution (Ballenberger *et al.*, 2012). However, it is not clear how to deal with censored data in the context of PCA, especially when there is a high fraction of censored data points. In this case deletion can result in the loss of an unacceptably high proportion of the data. Similarly, treating the censored data as missing (Theis *et al.*, 2011) discards potentially valuable information. Substitution can yield strongly biased results and multiple imputation results in several data-sets which are difficult to combine in a single PCA (Lubin *et al.*, 2004; Uh *et al.*, 2008). Furthermore, for very high fractions of censored data (as in single-cell qPCR), it is not clear how to derive adequate probability distributions (Ballenberger *et al.*, 2012; Lubin *et al.*, 2004).

When performing PCA of censored data from single-cell qPCR, the standard approach is to substitute the Ct values of all censored data-points with the same Ct value (usually LOD Ct) and perform standard PCA. In figure 2 a toy example is used to illustrate the issues of this substitution approach. A different approach is to treat the data as censored when performing the PCA; however, to date no algorithm allowing for linear and non-linear PCA of censored data has been presented. In the following we propose an extension of generalised PCA (Gaussian process latent variable models) allowing for censored data. When optimising the generative mapping (including the positions of the data points in a low-dimensional latent-space) we use statistically sound methods to account for the censoring such that uncertainties in the high-dimensional space are reflected in the low-dimensional visualisation of the data.

We evaluate our new strategy on dealing with single-cell qPCR data on two typical data-sets.

Thus, our contribution in this work is two-fold. First, we propose a strategy for performing PCA, and in fact probabilistic kernel PCA in general, of censored data. This allows the visualisation of censored data within the commonly used framework of PCA without introducing bias due to censoring and can be used for data from a wide range of sources. Second, we present a framework on how to account for uncertainties when performing PCA of single-cell qPCR data. We quantify potential new biological insights which can be gained by accounting for censoring: in the case of single-cell qPCR data, our approach can result in PCA representations which better reflect the underlying structure of the data and allow for a better identification of biologically meaningful sub-populations.

2 METHODS

In order to derive an algorithm for PCA of censored data, we first review probabilistic dual PCA before we show how we can use this as mathematical framework to deal with censored data.

2.1 Dual PCA for censored data

Standard PCA with Gaussian noise

Let the gene expressions in the data space be denoted by $Y = [y_1, \ldots, y_N]^T$, $y_i \in R^D$ and latent variables in the low-dimensional latent space be denoted by $X = [x_1, \ldots, x_N]^T$, $x_i \in R^Q$, with D being the dimension of the data

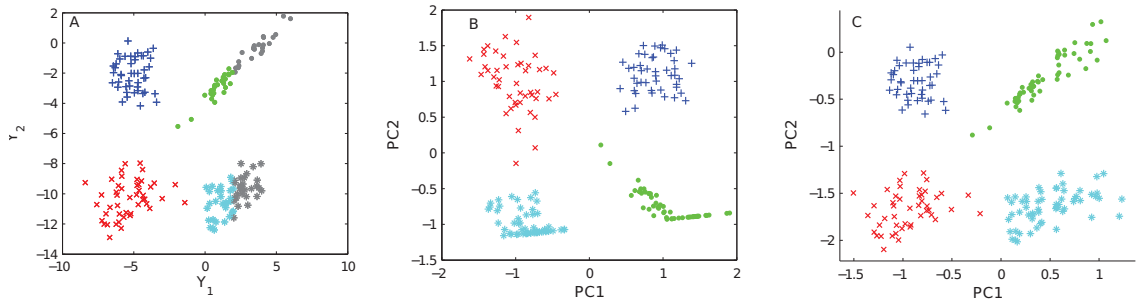


Fig. 2. 2D toy example (mixture of 4 Gaussians). In A the true values of Y are shown; Y_1 is right-censored for values greater than 2 (shown in gray). In B a PCA is performed with all censored values substituted with 2 resulting in a biased representation of the data. In C a PCA taking censoring into account using an appropriate noise model is shown resulting in a more realistic representation of the data. The uncertainty inherent in the generative model is visualised using grayscale as described in section 2.2. This uncertainty is greatest on the far right where censoring occurs.

space, Q the dimension of the latent space (usually 2 or 3) and N the number of samples in the dataset. Then, probabilistic PCA can be written as

$$y_n = W x_n + \eta_n \tag{1}$$

with i.i.d. Gaussian observation noise η_n: $p(\eta_n) = N(\eta_n|0, \beta^{-1}I)$ (Bishop, 2006). While for probabilistic PCA we would marginalise over X and optimise the transformation matrix W, for dual PCA (and more generally, GPLVM), we marginalise over W and optimise the latent variables X. If we place a prior over W in the form of $p(W) = \prod_{i=1}^{D} N(w_i|0, \alpha^{-1}I)$ where w_i is the ith row of W and integrate over W we find (Lawrence, 2004):

$$p(Y|X, \beta) = \frac{1}{(2\pi)^{\frac{DN}{2}}|K|^{D/2}} \exp\left(-\frac{1}{2}\mathrm{tr}(K^{-1}YY^T)\right) \tag{2}$$

with $K = \alpha XX^T + \beta^{-1}I$. This marginalised likelihood is the product of D Gaussian processes with linear covariance matrix K. It can be shown by deriving the corresponding log-likelihood L with respect to the latent variables X, the solution is equivalent to the one obtained by solving the standard PCA problem (Lawrence, 2005). In this dual interpretation of PCA, the cell-to-cell correlation is captured by the covariance matrix K. If the linear kernel in K is substituted with a different, non-linear kernel, a non-linear generalisation of probabilistic dual PCA (GPLVM) is obtained. By constructing the covariance matrix using such non-linear kernel, the relationship between cells can be arbitrarily complex. We chose the commonly used rbf kernel which can be written as:

$$k(x_1, x_2) = \alpha \exp\left(-\gamma(x_1 - x_2)^2\right) + \beta^{-1} \tag{3}$$

with hyperparameters α and γ.

Dual PCA with alternative noise models

So far the model assumes Gaussian noise η_n in every dimension which is a good approach when there are neither missing nor censored data. However, if we want to perform a (dual) PCA (or GPLVM) of censored or missing data, it is necessary to use a different noise model. This can be done by introducing an additional latent variable $F = [f_1, \ldots, f_N]$ between X and Y (Lawrence, 2005):

$$p(Y|X, \theta) = \int p(Y|F)p(F|X, \theta)dF = \int \prod_{n=1}^{N} p(y_n|f_n)p(F|X, \theta)dF \tag{4}$$

The Gaussian observation noise model used for non-censored data can then be interpreted as:

$$p(y_n|f_n) = \prod_{i=1}^{D} N(y_{ni}|f_{ni}, \beta^{-1}) \tag{5}$$

Lawrence (2005) suggested that other noise models in the form of

$$p(y_n|f_n) = \prod_{i=1}^{D} p(y_{ni}|f_{ni}) \tag{6}$$

can be used. However, in the case of non-Gaussian noise-models a Gaussian approximations of $p(y_{ni}|f_{ni})$ needs to be found in order to yield a Gaussian distribution of the posterior of F and thus maintaining the tractability of the marginalised likelihood.

$$p(y_{ni}|f_{ni}) \approx N(m_{ni}|f_{ni}, \beta_{ni}^{-1}) \tag{7}$$

Thus, in order to perform PCA/GPLVM with missing and censored data we first need to define an appropriate noise model. Next, once the (non-Gaussian) noise model is defined, we need to find a Gaussian approximation. Using this framework to deal with missing data is straight-forward: As Lawrence (2005) shows the precision β_{ni} corresponding to missing values ni is set to 0: $\beta_{ni} = 0$.

When dealing with censored data we need to define a more complex noise model. Here we propose to define a noise model based on the probit function (cumulative distribution function of the normal distribution)

$$\Phi(z) = \frac{1}{2\pi}\int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt.$$

For data points n which are right-censored at the value b in dimension i, a noise model reflecting this censoring can be defined as:

$$p(y_{ni} \geq b|f_{ni}) = \Phi(\lambda(f_{ni} - b)) \tag{8}$$

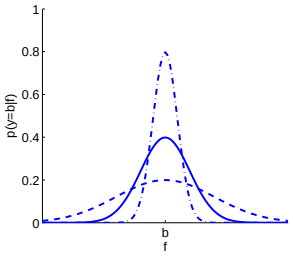
where λ controls the slope of the curve. For data which are left-censored at b_l in dimension i, we can choose the noise model accordingly ($p(y_{ni} \leq b|f_{ni}) = \Phi(-\lambda(f_{ni} - b_l))$). Similarly, data which are interval-censored between b_1 and b_2 can be accounted for with a noise model in form of

$$p(b_1 \leq y_{ni} \leq b_2|f_{ni}) = \Phi(\lambda(f_{ni} - b_1)) - \Phi(\lambda(f_{ni} - b_2)) \tag{9}$$

In figure 3 the probit noise model for $p(y \geq b|f)$ and the Gaussian noise model for $p(y = b|f)$ are shown. In the probit noise model the slope/steepness of the curve is controlled by the parameter λ; similarly, the width of the Gaussian noise model is controlled by β^{-1}.

Gaussian approximations as needed in equation (7) can be found by using assumed density filtering (ADF) (Minka, 2001; Lawrence *et al.*, 2005). Here, approximations are updated sequentially by incorporating one datum at a time. This yields an approximation $q(F)$ to the true posterior $p(F|X, Y)$ in the form of:

$$q(F) = N(f|f, \Sigma) \tag{10}$$



(a) Probit noise model (b) Gaussian noise model

Fig. 3. Probit noise model for 3 different vales of λ (a) and Gaussian noise model for 3 different values of β^{-1} (b).

with a block-diagonal covariance matrix Σ which is built of D blocks $\Sigma_1, \ldots, \Sigma_D$. The parameters of the approximation can be calculated as

$$\beta_{ni} = \frac{\nu_{ni}}{1 - \nu_{ni}\varsigma_{ni}} \tag{11}$$

$$m_{ni} = \frac{g_{ni}}{\nu_{ni}} + f_{ni} \tag{12}$$

with ς_{ni} being the nth diagonal element of Σ_i, $g_{ni} = \frac{\partial}{\partial f_{ni}} \ln Z_{ni}$ and $\nu_{ni} = g_{ni}^2 - 2\frac{\partial}{\partial \varsigma_{ni}} \ln Z_{ni}$. The partition function Z_{ni} is defined as

$$Z_{ni} = \int p(y_{ni}|f_{ni})q(F)dF \tag{13}$$

It can be shown (Lawrence *et al.*, 2005) that for the case of the probit noise model the partition function can be calculated as:

$$Z_{ni} = \Phi(u_{ni})$$

where

$$u_{ni} = c_{ni}(f_{ni} - b)$$

$$c_{ni} = \frac{1}{\lambda^{-2} + \varsigma_{ni}} \tag{14}$$

In practice, if the slope of the noise model is not fixed, we learn it together with the kernel parameters: therefore we consider the slope of the noise model to be very steep and add a white noise term to the kernel K in form of $\lambda^{-2}I_c$ with I_c being a diagonal matrix such that only the entries corresponding to censored data points are set to 1 and all other entries are set to 0 — this will then result in an increase of ς_{ni} by λ and, as can be seen from equation (14), in an equivalent description of the noise model. Note that care has to be taken as censored inputs are independent for each dimension. This means that we have to use a different kernel for each dimension as I_c will be different for each dimension. However, as the marginal likelihood factorises into d Gaussian processes this extension of standard GPLVMs is straightforward and the possibility was in fact described earlier (Grochow *et al.*, 2004). More specifically, we choose the white noise term for dimension d such that $k_{w,d} = \lambda^{-2}I_{c,d} + \sigma^2 I_{nc,d}$ with $I_{c,d}$ and $I_{nc,d}$ being the diagonal matrices where only those entries are set to 1 where a data-point is censored and not censored respectively. All other terms in the kernel (i.e. rbf term or linear term) were shared across all dimensions.

In summary, the generation of the PCA mapping taking censoring into account involves two major steps. First, a Gaussian approximation (eq. (7)) to the probit noise-model (eq. (8)) has to be found via ADF (eq. (11) and (12)). This yields an approximation to the log-likelihood of the model (eq. (4)). In a second step, this approximation is maximised with respect to the latent positions X and the kernel parameters (including λ). For this optimisation step non-linear optimisers such as scaled conjugate gradient (Nabney., 2001) can be used.

2.2 Visualising uncertainties in latent space

Performing dual PCA with the probit noise model as outlined above, yields an explicit mapping from latent space to the original high-dimensional space (eq. 4). When generating this mapping, not only the positions of the points in the latent space, but also the parameters of the (noise) model are chosen such that censoring is accounted for. That is, the uncertainty of the data is reflected in the mapping.

Consequently, we can use this model to calculate for any point x^* in the latent space a posterior mean $M_i(x^*)$ and a posterior variance $V_i(x^*)$ for each dimension i (Suppl. note 1) (Rasmussen and Williams, 2006). For standard GPLVMs with one kernel shared across all dimensions $V_i(x^*)$ will be the same for all dimensions. In this case it is straight-forward to visualise the uncertainty of the mapping in the latent space by varying the intensity of the background pixels (background of the 2D map) (Lawrence, 2006). In our case the posterior variance will vary across dimensions. In order to visualise the uncertainty across all dimensions, we use the fact that eq. (4) is a product of D Gaussian processes. Consequently, we can quantify the uncertainty of the mapping by calculating the product of the the posterior variance across all dimensions. For visualising this uncertainty we then vary the intensity of the background pixels with

$$\text{Intensity}(x^*) \propto \sum_{i=1}^{D} \log(V_i(x^*)) \tag{15}$$

The higher the uncertainty, the darker the pixels. Black pixels correspond to the highest uncertainty.

2.3 PCA of censored single-cell qPCR data

Censoring in single-cell qPCR techniques occurs due to a detection limit of the qPCR reaction. This limit of detection both depends on the manufacturer of the machine and is experiment-specific where it can vary between different genes. Most researchers do not establish this gene-dependent LOD but use a global LOD reflecting the overall sensitivity of the qPCR machine (Guo *et al.*, 2010; Pina *et al.*, 2012). However, more objective methods to establish an LOD such that a qPCR reaction will be found with a probability of at least 99% if the Ct value is below the LOD, can be used, too. For example, necessary experiments to do so are outlined in the manual of the popular Biomarks system (Fluidigm Corporation, 2012).

In the following we will evaluate our new strategy on how to deal with the limit of detection (once it is established) when performing PCA. Therefore, we first use the standard approach where all values greater than the limit of detection are substituted with a particular value Ct_{sub}. The choice of Ct_{sub} depends largely on the biological interpretation of non-detects (Suppl. Note 2). If most non-detects correspond to a genuine lack of transcription, a large value should be chosen for Ct_{sub} as the true underlying Ct value would be ∞ (for practical reasons $\text{Ct}_{\text{sub}} = 40$ could be chosen, as maximum of 40 cycles can typically be measured); otherwise a value of Ct_{sub} closer to LOD (or LOD) should be chosen. We followed the latter approach (setting Ct_{sub} =LOD) which is commonly adopted in the literature (Pina *et al.*, 2012; Guo *et al.*, 2010). Furthermore we also explored higher values of Ct_{sub}, corresponding to the interpretation of non-detects as lack of transcription (Supp. Fig. 1-2). As in the substitution approach systematic uncertainties in the data in form of censoring are ignored, in this case standard PCA can be performed. In addition to standard PCA, results for ICA and t-SNE (van der Maaten and Hinton, 2008; Amir *et al.*, 2013) using the substitution approach are shown in Suppl. Fig. 1-2. We then compare this substitution approach to our new algorithm where PCA with the probit noise model is performed. In contrast to the substitution approach there is no need to choose Ct_{sub} as the probit noise model accounts for uncertainties in the underlying true Ct value for non-detects. As the non-detects are modelled separately

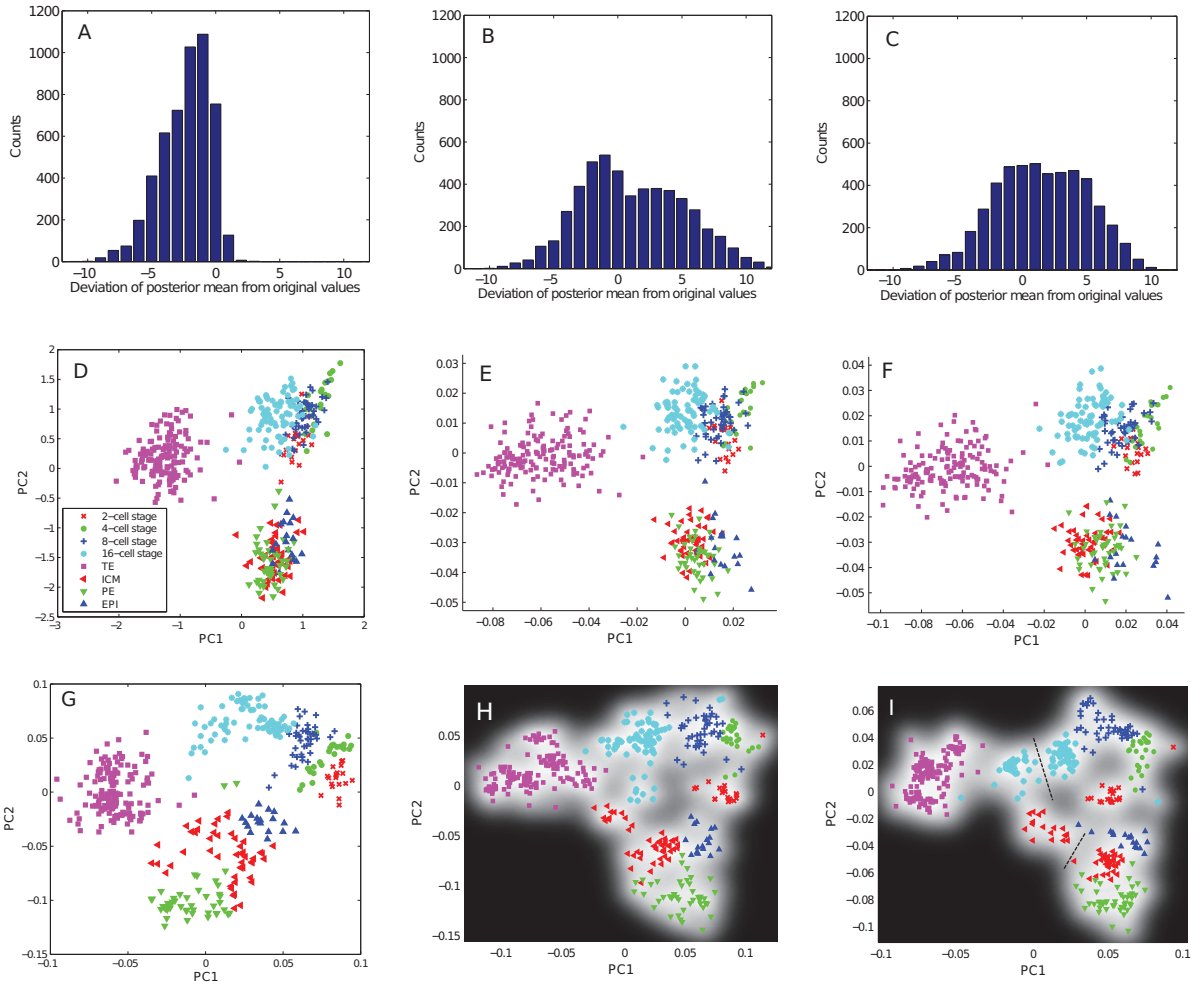


Fig. 4. A to C: Distribution of residuals between posterior means and the normalised LODs for differnt approaches. D to F:PCA with censored data from mESC data-set. Standard PCA with substitution approach (D), taking censoring into account with probit noise model and fixed λ (E) and probit noise model with λ learnt from data (F). G to J: GPLVM with rbf kernel for mESC data. Standard GPLVM with substitution approach (G), taking censoring into account with probit noise model and fixed λ (H) and probit noise model with λ learnt from data (I). In (I) the dashed lines indicate two distinct subpopulations at the 16-cell stage and ICM.

by introducing a discrete part in the GPLVM, this can either be interpreted as a noise model for censored data or as a discrete model for genes which are "off". For the probit noise model we compare censored PCA with fixed steepness parameter λ to censored PCA where λ is optimised in form of a parameter of a white noise kernel as described above.

We tested the two approaches with both a linear kernel (resulting in standard PCA) and an rbf kernel in order to capture non-linearities in the data.

As theoretically a maximum of 40 cycles can be measured, we used this as upper limit in the noise model for interval censoring (equation (9)). This prevents the optimiser from being stuck in a local minimum where some censored data-points are mapped to very high Ct numbers[2]. In each run we followed Guo *et al.* (2010) and Moignard *et al.* (2013) and performed a cell-wise normalisation by subtracting the average Ct number of the housekeeping genes from the Ct value of the gene of interest. Consequently, when data-points where censored at a value b before normalisation, this threshold was normalised accordingly for each cell (Ballenberger *et al.*, 2012).

We evaluated the different approaches on two recently published data-sets. The first data-set was published by Guo *et al.* (2010). Briefly, the authors analysed the development of the mouse zygote to the bastocyst by measuring gene expression on a single-cell level. Therefore, the authors quantified the expression levels of 48 genes for a total of 442 cells at different stages of the cellular development (one-cell stage to 64-cell stage). Cells at the 32-cell stage had undergone differentiation to either trophoectoderm (TE) cells or inner cell mass (ICM). Cells at the 64-cell stage were either TE cells, primitive endoderm (PE) cells or epiblast (EPI) cells. Lables for cells at 32-cell stage and 64-cell stage were derived from figure 1 in (Guo *et al.*, 2010) by assigning each cell to the closest cluster (TE, PE, EPI, ICM). Cells from the 1-cell stage were systematically different from all other cells due to differences in experimental conditions (Guo *et al.*, 2010). That is why we excluded all 9 cells from the one-cell stage from our analysis. More details on the data-set can be found in recent publications by Guo *et al.* (2010) and Buettner and Theis (2012).

[2] In practice this only occurred for the linear kernel with fixed λ in the blood data-set.

The second data set consists of 597 blood stem and progenitor cells, in which the expression of 24 genes was measured, including 18 transcription factors, five housekeeping genes and a cell surface marker Moignard *et al.* (2013). Approximately 120 individual primary cells were isolated for each population from mouse bone marrow by fluorescence activated cell sorting (FACS). As for the ESC data, the sorted populations comprise a cellular hierarchy which gives rise to all of the mature cell types of the blood system. The haematopoietic stem cells sit atop the hierarchy and give rise to the megakaryocyte-erythroid lineage through the PreMegE progenitor, or to the lymphoid-primed multipotent progenitor (LMPP). The LMPP in turn gives rise to the myeloid lineage and the lymphoid lineage through the granulocyte-monocyte progenitor (GMP) and the common lymphoid progenitor (CLP), respectively (Orkin and Zon, 2008). Each population has been isolated on the basis of cell surface markers and characterized functionally either in vivo or in vitro.

In the primary analysis of both data-sets a limit of detection of Ct=28 was assumed.

For both datasets we evaluated our new approach to deal with censoring for both a linear and an rbf kernel. First, we assess the effect of the probit noise model compared to the Gaussian noise model. Therefore we make use of the generative models and calculate the posterior mean of $p(y_c|x_c)$ for all censored data points c. Furthermore we quantified the performance of the different approaches in terms of their ability to reflect the known structure of the data by calculating the nearest neighbour error: for each cell we established the label of its nearest neighbour in the respective 2D space; if the label differed from the original cell, we increased the nearest neighbour error count by one. We chose this metric as it is easily interpretable and commonly used in the machine learning community to quantify the performance of dimensionality reduction/visualisation methods; however, as its power is limited (e.g. it does not account for newly discovered sub-populations), visual inspection as additional performance measure is crucial.

3 RESULTS

In figure 1 the fraction of censored data-points in both data-sets is illustrated for the different genes. It can be seen that in both data-sets a considerable fraction of data is censored across some dimensions (genes), while for other dimensions no censoring occurred (i.e. expression of the respective gene could be detected with a Ct number below the LOD for all cells.)

In the following we will first evaluate different approaches of PCA with censored data for the mESC data. Next, we will repeat the evaluation with a different dataset on blood stem/progenitor cells.

3.1 Evaluation of censored PCA for mESC data

The result of a standard PCA where all censored values are substituted with LOD (as described in section 2.3) is shown in figure 4(D). This method was used in the original publication and yielded a nearest neighbour error of 124. In the original high-dimensional data-space the nearest neighbour error was 10. Note that this error was calculated using the substitution approach.

While TE cells can be clearly distinguished from all other cell types, early cells from 2-cell stage to 8-cell stage are strongly overlapping. Similarly, there is a strong overlap between ICM cells and PE/EPI cells.

Next, we compare the substitution method with our new algorithm for censored PCA. In figure 4(E) and (F) the results for a fixed $\lambda^{-2} = 10$ are shown together with the representation where λ was optimised together with the other kernel parameters.

For a quantitative analysis of the effects of the different noise models, we used the generative mapping from the latent space to the high-dimensional space to calculate the posterior means given all censored data-points in the low-dimensional space. We then calculated the residuals between the posterior means and the respective normalised LOD. In figure 4 (A) to (C) it can be seen that when using the substitution approach, the censored values are mapped consistently to values lower than the normalised LOD. In contrast, when our new approach is used, a large fraction of censored data-points is mapped to values greater than the normalised LOD, which is in better agreement with the ground truth. When λ was learnt from the data by optimising it in form of a kernel parameter, the maximum a posteriori estimate was 15.3 for λ^{-2}. Compared to a fixed value for λ^{-2} of 10, censored data-points were mapped closer to the normalised LOD. It can be seen that taking into account the censoring results in an improved mapping where EPI cells can be separated better from ICM/PE cells than in the standard method. This is reflected in lower nearest-neighbour errors of 113 and 88 for fixed λ and learnt λ respectively.

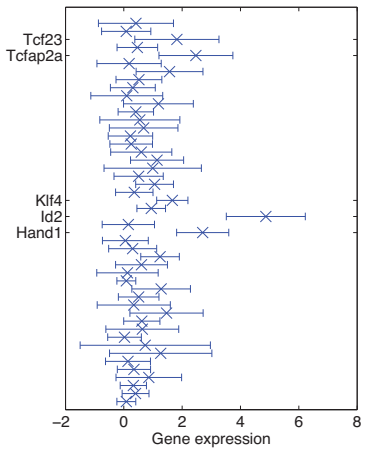


Fig. 5. Difference in gene expression between the 2 subclusters at the 16-cell stage for different mappings. The error bars show the variation of gene expression within the smaller sub-cluster (1 standard deviation in each direction). For convenience, genes with the greatest differences are labelled in the plots.

We also evaluated our new approach for an rbf kernel, which allows non-linearities to be taken into account. The resulting mappings are shown in figure 4 (G) to (I).

It can be seen that in the non-linear case, the separation between different time-points and cell-times is comparable between the substitution approach and our new approach. This is also reflected in the similar nearest neighbour errors of 11, 12 and 10 respectively.

However, it can be seen that the ICM cells as well as cells from the 16-cell stage are separated into two clusters when the censoring is accounted for. This leaves room for interpretation. When comparing mean gene expression for the two sub-clusters in the 16-cell stage we found that expression in Id2 and Klf4 differed considerably between the two subclusters (p-values after Wilcoxon rank sum test 10^{-6} and 10^{-11} respectively, figure 5). This is in good agreement

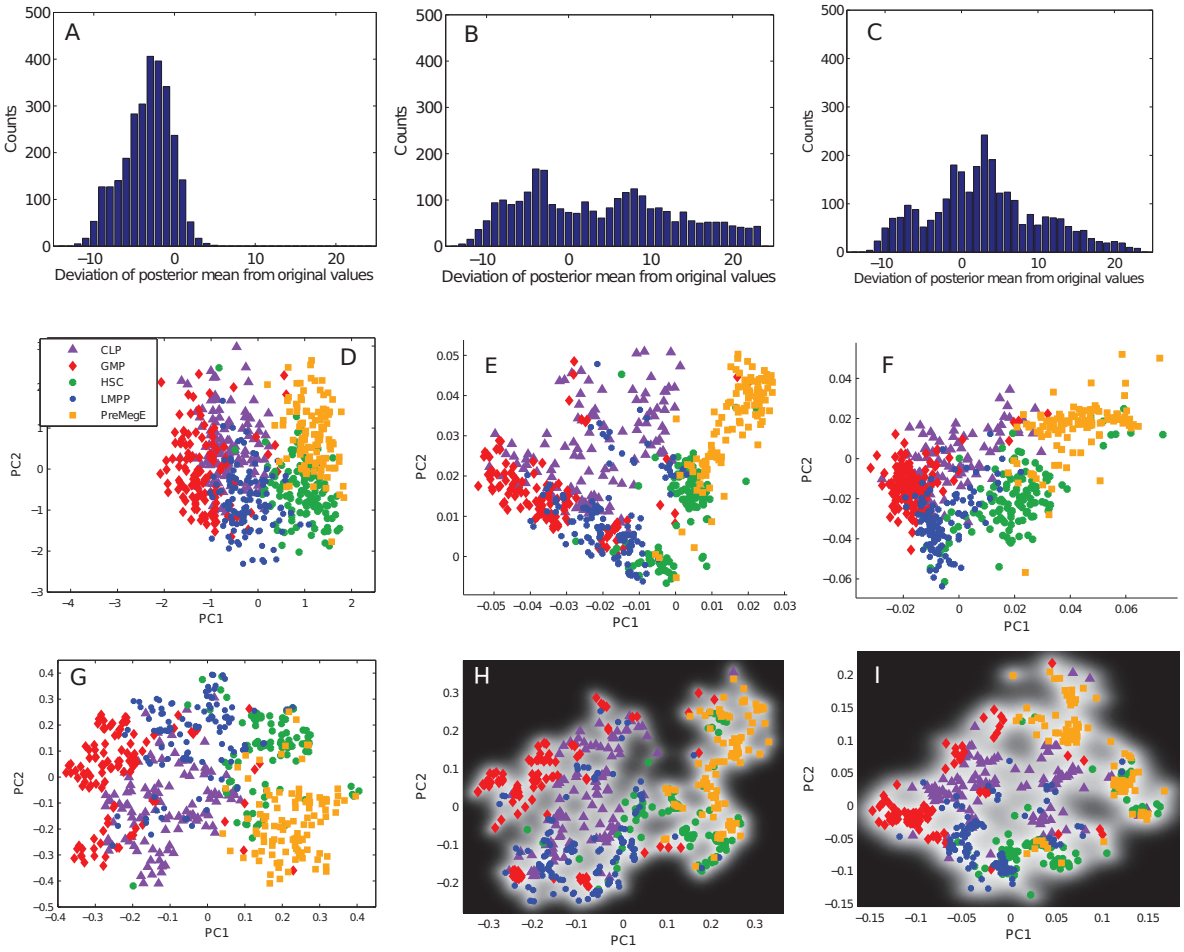


Fig. 6. A to C: Distribution of residuals between posterior means and the normalised LODs for differnt approaches. D to F: PCA with censored data from blood data-set. Standard PCA with substitution approach (D), taking censoring into account with probit noise model and fixed λ (E) and probit noise model with λ learnt from data (F). G to J: GPLVM with rbf kernel for blood data. Standard GPLVM with substitution approach (G), taking censoring into account with probit noise model and fixed λ (H) and probit noise model with λ learnt from data (I). The background intensity indicates the relative uncertainty of the mapping with black pixels corresponding to the highest uncertainty of the mapping[3].

with previously reported experimental results from Guo *et al.* (2010) who show that Id2 is the earliest markers for outer cells. Similarly, when comparing mean gene expressions in the 2 subclusters of ICM cells we found that they differed significantly in expression of Fgf4 (p=0.01, Wilcoxon rank sum test). This is also in good agreement with previously reported results showing differential expression of Fgf4 in the early inner cell mass Guo *et al.* (2010). Thus, when allowing for non-linearities and taking censoring into account, it was possible to correctly represent the structure of the data for all cell-types and resolve subpopulations which could not be revealed when not accounting for censoring. In supplementary figure 3 the nearest neighbour errors for all approaches to perform a PCA of the mESC data-set are shown. In supplementary figure 1 we show results for the substitution approach with other multivariate methods for different choices of Ct_{sub}. All approaches yielded higher nearest neighbour errors than the GPLVM with probit noise model.

3.2 Evaluation of censored PCA for blood stem/progenitor cell data

To evaluate the potential benefits of our new approach for PCA of censored data with a second independent biological dataset, we next applied our new analysis tools to a recently generated single cell gene expression dataset for 5 FACS sorted populations of blood stem and progenitor cells.

As for the mESC data set we first compared standard PCA with the substitution approach to censored PCA with the probit noise model. Results are shown in figure 6 (D) to (F). It can be seen that by accounting for censoring in the data, a better separation is achieved between most cell types; this occurs most clearly for CLPs and GMPs. Consequently nearest neighbour errors decreased from 254 errors with the standard substitution approach to 193 and 217 when censoring was accounted for by fixing λ and learning λ respectively. As for the mESC data-set we found that the censored PCA approach yielded better posterior mean values for censored

data-points than the standard approach using substitution (figure 6 (A) to (C)).

We also used an rbf kernel to evaluate the non-linear PCA with censored data. Results for the different approaches are shown in figure 6 (G) to (I). When accounting for the censoring, the nearest neighbour error was reduced and a better separation between CLPs and LMPPs than for the substitution approach was possible. Nearest neighbour error rates for all approaches are summarised in supplementary figure 3. In supplementary figure 2 we show results for the substitution approach with other multivariate methods for different choices of Ct_{sub}. All approaches yielded higher nearest neighbour errors than the GPLVM with probit noise model.

4 DISCUSSION

Conventional approaches for PCA of censored data where values beyond the detection limit are substituted with the detection limit can yield strongly biased results. We have proposed a novel approach for performing dual PCA of censored data. Our new approach resulted in a mapping between low dimensional and high-dimensional space such that more censored data-points were mapped correctly to values greater than the detection limit. It was previously shown that for single-cell qPCR data it is crucial to explicitly model the population of non-detects when performing a statistical test of univariate differential expression (McDavid *et al.*, 2013). To date no approaches for dealing with this issue for multivariate analyses such as PCA have been proposed. We evaluated our new approach for two different real-world data-sets comprising measurements of single-cell qPCR data. For both data-sets the PCA representations better reflected the known structure of the data when the censoring was explicitly considered. We evaluated using a linear as well as a non-linear kernel and for both data-sets accounting for non-linearities resulted in better visualisations. In contrast to using a linear kernel (i.e. PCA), this comes at the price of losing interpretability - while in the linear case loadings can be easily visualised in a bi-plot, in the non-linear case this is more difficult as loadings change across the 2D plot. Whether trading off interpretability for complexity is beneficial depends highly on the data-set under consideration and any non-linearities present. In the context of single-cell qPCR data our analyses suggest that a non-linear kernel is necessary to capture the typical complex dependency structure of such data.

For linear kernels (corresponding to standard PCA) as well as for non-linear kernels (allowing for interactions), our new approach yielded considerably lower nearest-neighbour error rates with reductions of up to 29% in the linear case. Furthermore, in the case of mESC data, the structure of sub-populations was reflected better in the case when censoring was taken into account in the non-linear case: in contrast to non-linear probabilistic PCA with the substitution approach, two sub-populations corresponding to cells from the 16-cell stage with high Id2 expression and cells in the inner cell-mass with high Fgf4 expression could be identified. These known subpopulations were previously identified in univariate analysis of cells from the same cell stage. However, this standard approach has several drawbacks as it can become unfeasible when too many genes are measured simultaneously. Furthermore, only univariate patterns can be identified, while important information may lie in multivariate patterns which could be defined by the differential expression of several genes. Finally, when analysing univariate distributions or correlations between two genes for cells from the same cell-stage, the identified subpopulations cannot be put in context with other cells from other cell-stages. In contrast, when performing a probabilistic (kernel) PCA of all cell stages, it is possible to identify complex multivariate subpopulations and by simultaneously displaying all cells, the PCA plot provides an intuitive illustration of the relation between all cell-populations.
This was achieved by implementing a Gaussian-process latent variable model with different kernels for each dimension. Censoring was accounted for by a probit noise level. The steepness parameter of the probit function was learnt together with other kernel parameters, resulting in a parameter-free approach for PCA of censored data.

While our approach was designed for accounting for uncertainties in single-cell qPCR data, related issues can be found in single-cell RNAseq data. In contrast to single-cell qPCR however, very high levels of technical noise are present in all commonly used protocols for single-cell RNAseq (Brennecke *et al.*, 2013). This technical noise is particularly strong for low levels of expression and dominates all other uncertainties (like censoring). Although these uncertainties are inherently different from the censoring found in single-cell qPCR, the flexible framework of Gaussian Processes allows to account for these uncertainties in a straight forward manner by using and additional term in the (Gaussian) noise model reflecting the technical noise which can be estimated using the approach suggested by Brennecke *et al.* (2013). While single-cell RNA-Seq data is generated in form of read counts, it is crucial to perform normalisation steps accounting for different cell sizes, different sequencing depth and, depending on the protocol, different transcript lengths (Yan *et al.*, 2013; Brennecke *et al.*, 2013). Such normalisation can be performed by calculating RPKM, FPKM or using DEseq inspired normalisation procedures (Anders and Huber, 2010; Brennecke *et al.*, 2013). After normalisation, gene expression is measured on a continuous scale such that - after an appropriate variance-stabilising transformation (e.g. log-transformation) - GPLVM can be applied without modification. As efficient implementations allow fast processing of datasets with tens of thousands of genes and hundreds of cells without overfitting, it is a promising tool for analysing such datasets.

The main drawback of our proposed approach is that it scales cubically with the number of cells with which may be prohibitive when the number of analysed cells is very large ($>> 10^4$). While standard GPLVMs are time-consuming, too, significant speed-ups can be achieved due to sharing the kernel across all dimensions and using a spherical noise model. However, if necessary, approximations resulting in sparse covariance matrices commonly used in Gaussian process literature, could be applied for our framework too. For the application to single-cell qPCR data, we found that this was not necessary as computation times were in the order of only few hours on a standard laptop. We acknowledge that this is a considerable increase of time compared to standard PCA which can be performed when using the substitution approach to deal with censored data. In applications with only a small fraction of censored data-points this rather large increase in runtime may only result in minor changes of the PCA representation and simpler approaches such as the substitution approach or treating the data as

missing may be a valid alternative if runtime is an issue. However, in the case of single-cell qPCR data we have shown that taking censoring into account avoids a potential bias in low-dimensional representations due to the censoring. This in turn can result in better biological insights: first, our approach can yield a better separation of different cell types and second, even reveal new biologically meaningful sub-populations which may be obscured due to a bias introduced by the censoring. When designing single-cell qPCR experiments, the quantification of heterogeneities and the reliable identification of new sub-populations are often key goals. That is why we believe that our approach will be of interest for many practitioners working with censored data, especially in the field of high-throughput single-cell qPCR.

5 CONCLUSION

We have presented a new approach for performing probabilistic PCA for censored data within the framework of Gaussian process latent variable models. Therefore we implemented an appropriate noise model and allowed different kernels for each dimension. We showed that for single-cell qPCR data with a high fraction of censored data-points the resulting probabilistic (kernel) PCA representations reflected the true structure of the data better than conventional approaches. In two real world datasets known cell types could be better separated when censoring was taken into account and in one dataset several distinct subpopulations could be revealed which could not be resolved with standard PCA.

Funding: FB and FJT acknowledge funding from the European Research Council (starting grant "LatentCauses"). VM and BG acknowledge funding from Leukaemia and Lymphoma Research (LLR), Cancer Research UK (CRUK) and the Biotechnology and Biological Sciences Research Council (BBSRC). VM is funded by a Medical Research Council studentship.

REFERENCES

Amir, E.-a. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*, **31**(6), 545–552.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106.

Ballenberger, N., Lluis, A., von Mutius, E., Illi, S., and Schaub, B. (2012). Novel statistical approaches for non-normal censored immunological data: analysis of cytokine and gene expression data. *PLoS One*, **7**(10), e46423.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., *et al.* (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*.

Buettner, F. and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, **28**(18), i626–i632.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., and Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol*, **29**(12), 1120–1127.

Dominguez, M. H., Chattopadhyay, P. K., Ma, S., Lamoreaux, L., McDavid, A., Finak, G., Gottardo, R., Koup, R. A., and Roederer, M. (2013). Highly multiplexed quantitation of gene expression on single cells. *J Immunol Methods*, **391**(1-2), 133–145.

Fluidigm Corporation (2012). Application guidance: Single-cell data analysis.

Grochow, K., Martin, S. L., Hertzmann, A., and Popovic., Z. (2004). Style-based inverse kinematics. In *ACM Transactions on Graphics (SIGGRAPH 2004)*.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell*, **18**(4), 675–685.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, **6**, 1783–1816.

Lawrence, N., Platt, J., and Jordan, M. (2005). Extensions of the informative vector machine. In J. Winkler, M. Niranjan, and N. Lawrence, editors, *Deterministic and Statistical Methods in Machine Learning*, volume 3635 of *Lecture Notes in Computer Science*, pages 56–87. Springer Berlin Heidelberg.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In *In NIPS*, page 2004.

Lawrence, N. D. (2006). Local distance preservation in the gp-lvm through back constraints. In *In ICML*, pages 513–520. ACM Press.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect*, **112**(17), 1691–1696.

McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, **29**(4), 461–467.

Minka, T. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA. Morgan Kaufmann.

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., Jacobsen, S. E., de Bruijn, M. F., and Göttgens, B. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol*, **15**(4), 363–372.

Nabney., I. T. (2001). Netlab: Algorithms for pattern recognition. In *Advances in Pattern Recognition*. Springer, Berlin.

Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, **132**(4), 631–644.

Pina, C., Fugazza, C., Tipping, A. J., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. *Nat Cell Biol*, **14**(3), 287–294.

Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press.

Taniguchi, K., Kajiyama, T., and Kambara, H. (2009). Quantitative analysis of gene expression in a single cell by qpcr. *Nat Methods.*, **6**(7), 503–506.

Theis, F. J., Latif, N., Wong, P., and Frishman, D. (2011). Complex principal component and correlation structure of 16 yeast genomic variables. *Mol Biol Evol*, **28**(9), 2501–2512.

Uh, H.-W., Hartgers, F. C., Yazdanbakhsh, M., and Houwing-Duistermaat, J. J. (2008). Evaluation of regression methods when immunological measurements are constrained by detection limits. *BMC Immunol*, **9**, 59.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., and Tang, F. (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol*, **20**(9), 1131–1139.